Noninvasive selective activation and suppression of the human brain with natural images

by

Andrii Zahorodnii

B.S. in Computation and Cognition, MIT, 2024

Submitted to the Department of Brain and Cognitive Sciences and
the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTATION AND COGNITION

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2026

© 2026 Andrii Zahorodnii. All rights reserved.

The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by: Andrii Zahorodnii
Department of Brain and Cognitive Sciences
Department of Electrical Engineering and Computer Science
February 3, 2026

Certified by: Ila Fiete
Associate Investigator, McGovern Institute
Professor, Brain and Cognitive Sciences, MIT;
Director, K. Lisa Yang ICoN Center, MIT, Thesis Supervisor

Accepted by: Mehrdad Jazayeri
Investigator, McGovern Institute
Professor and Director of Education, Brain and Cognitive Sciences, MIT;
Investigator, Howard Hughes Medical Institute

Noninvasive selective activation and suppression of the human brain with natural images

by

Andrii Zahorodnii

Submitted to the Department of Brain and Cognitive Sciences and
the Department of Electrical Engineering and Computer Science
on February 3, 2026 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTATION AND COGNITION

ABSTRACT

The ability to selectively modulate neural activity in the human brain has profound implications for both basic neuroscience and clinical applications. However, approaches for achieving neural control in humans remain limited. Here we demonstrate that deep neural network models trained on intracranial recordings can predict which natural images will drive or suppress activity at specific cortical sites in the human brain. We recorded stereotactic electroencephalography (SEEG) from epilepsy patients viewing thousands of images in a rapid serial visual presentation experiment, establishing the first human intracranial dataset designed for model-based stimulus optimization. Using pretrained image models combined with linear regression, we identified visual stimuli that successfully modulated neural activity in predicted directions at targeted recording sites. We found that visual reliability was highest in occipital and temporal regions, though neural target stability varied substantially across recording days. Our systematic analysis of scaling laws revealed that model performance improves approximately linearly with the logarithm of training set size, with stimulus diversity contributing more than repeated presentations for a fixed experimental budget. Models improved iteratively when retrained with feedback from previous sessions, demonstrating successful closed-loop refinement of model predictions. Our framework extends prior work in non-human primates to human intracranial electrophysiology, establishing a proof-of-concept for stimulus-based neural control in humans. These results open new avenues for non-invasive therapeutic interventions for neurological disorders and for causal testing of computational theories of human sensory processing.

Thesis supervisor: Ila Fiete
Title: Associate Investigator, McGovern Institute
Professor, Brain and Cognitive Sciences, MIT;
Director, K. Lisa Yang ICoN Center, MIT

I would like to thank Dr. Joshua Aronson for inviting me to join his clinical team at Beth Israel Deaconess Medical Center, teaching me how to conduct SEEG experiments, supervising the entire process, guiding me through task design and analysis, and providing valuable feedback throughout the project.

I am deeply grateful to my advisor, Prof. Ila Fiete, for her guidance throughout this work and throughout my undergraduate and master's research. I am particularly grateful for her sharing research expertise and wisdom, and giving me the freedom to choose my own research direction and pursue it.

I would like to thank Riley DeHaan for his substantial help with the technical setup for neural data recording, for troubleshooting data collection issues, and for offering thoughtful feedback on the core ideas of this project in its earliest stages.

I am also thankful to Prof. James DiCarlo and Dr. Guy Gaziv for helpful discussions and their suggestions on analysis and task design early in this project. I would like to thank Dr. Yoon Bai for walking me through and answering all of my questions about the experimental setup in the DiCarlo lab, which I adapted to the hospital environment for this project. This project builds on top of DiCarlo lab's work in modeling the visual stream, and extends it to the human brain.

Finally, I wish to express my gratitude to the patients at BIDMC who participated in this research. Their willingness to contribute to science has enabled this project and helped advance the understanding of the human brain.

Contents

List of Figures 9

List of Tables 15

1 Introduction **17**

1.1 Related Work . . . 18

2 Results **21**

2.1 Visual reliability varies across brain regions . . . 21

2.2 Stability of neural targets over time . . . 23

2.3 Predicting neural responses from image features . . . 27

2.4 Driving and suppressing neural activity with model-selected stimuli . . . 27

2.5 Iterative model improvement with feedback data . . . 29

2.6 Scaling laws of neural prediction . . . 31

2.7 Interpreting model selectivity . . . 32

2.8 Selectivity of the stimulation protocol . . . 33

3 Methods **37**

3.1 Study Overview and Participants . . . 37

3.2 Recording Setup . . . 37

3.3 Task Design . . . 39

3.4 Visual Stimuli . . . 41

3.5 Signal Processing . . . 42

3.6 Split-Half Correlation Analysis . . . 42

3.7 Automatic Target Selection Pipeline . . . 43

3.8 Stability Analysis Over Time . . . 47

3.9 Predicting Neural Responses from Image Features . . . 48

3.10 Scaling Laws Analysis . . . 50

3.11 Driving and Suppressing Neural Activity . . . 50

3.12 Iterative Model Improvement . . . 51

3.13 Model Interpretation . . . 51

3.14 Software . . . 51

4 Discussion **55**

A Appendix: Stability of Targets Across Days **59**
A.1 Intra-day Reliability Remains Stable Across Recording Days 59
A.2 Raw Signal Distribution Analysis Across Recording Days 59
A.3 Relationship Between Distribution Stability and Cross-Day Image Stability . 63

B Appendix: Advanced, Temporally Precise Trigger Detection **65**
B.1 Hardware Configuration 65
B.2 Photodetector Response Characteristics 66
B.3 Trigger Detection Pipeline 66
B.4 Validation of Trigger Extraction 68

C Appendix: Extra Figures **71**

References 73

List of Figures

2.1 **Experimental setup, model, and key questions.** (a) Schematic of a subject with implanted SEEG probes viewing successive images on the screen. Neural data was recorded as subjects viewed images in a rapid serial visual presentation (RSVP) setup. (b) The RSVP task. Images from various datasets (OASIS, ImageNet, and TrailerFacesHQ) were presented in rapid succession (on average every 250 ms), with subjects required to maintain fixation on a central cross and perform an auxiliary attention task (see Methods). (c) Overview of our model-based experimental approach and main questions. On the left, a sketch shows a set of responses to training images being provided as input to a pre-trained neural network model (such as a convolutional neural network, depicted as layers), followed by a linear readout layer that is trained to predict the neural response recorded at a specific SEEG site. Once trained, this model enables us to take a large set of candidate images and predict which stimuli will most strongly drive or suppress the targeted neural site. 22

2.2 **Split-half correlation reveals neural sites with reliable visual encoding in the human cortex.** For results in all neural sites, see Figures 3.3, 3.4, 3.5. 24

2.3 **Visual reliability is highest in occipital and temporal cortex.** Orthogonal brain map showing all SEEG neural sites as dots. Dot size and color represent the split-half correlation coefficient R computed using one repeat per image. Highest reliability is observed in occipital and temporal regions. Based on the measure from Figure 2.2. 25

2.4 **Some neural sites maintain stable representations across days while others degrade.** Scatter plots for four example neural sites showing response correlations across days. For each target, the y-axis shows the neural response on the first recording day (Day 2). The x-axis shows: (left) the response to the same image presented again on the same day; (middle) the response on Day 5; (right) the response on Day 7 (subject MIT003) or Day 13 (subject MIT002). Each dot represents one image. Some targets like ROLI7-8 maintain high correlations across all days, demonstrating remarkable stability throughout the subject's hospital stay. Other targets like RTHA2-3 show substantially degraded stability over time. In this figure, each response was averaged across 4 repeats per image. 25

2.5 **A minority of neural sites are both reliable and stable across days.** Bar plot showing the proportion of neural sites in each experimental subject across three categories: total sites (all electrode pairs), reliable sites, and stable+reliable sites (see Main Text and Methods for definitions). Only a subset of all sites show reliable visual responses, and only a further subset of those maintain stability across recording days. 26
2.6 **Neural target stability varies over time post-surgery.** Y-axis shows stability, defined as the correlation of neural responses for the same set of images presented on different recording days, divided by the split-half correlation on the first recording day (at 4 repeats per image). All stability values are further normalized by dividing by their value on the first recording day. X-axis shows the post-surgery day. Only reliable targets are included in this analysis. . . 26
2.7 **Adversarially trained ResNet-50 achieves highest prediction of neural responses.** Left: Test R^2 for four example neural sites across different pretrained architectures. Right: Mean test R^2 across $N = 23$ stable and reliable targets, normalized by the noise ceiling of each target. Each bar corresponds to a different pretrained image model. Training and test sessions were conducted on different days, providing a strong test of model generalization and ensuring that models capture stable aspects of neural representations. All test data consisted of 4 repeats per image, regardless of the training setup. Error bars represent standard error of the mean across targets. 28
2.8 **Model-selected stimuli successfully drive and suppress neural activity.** Each panel shows one neural target. Three bars represent mean voltage in the target neural site for different image categories: (gray) random images – randomly selected new images shown on that day ($N > 200$); (red) drive images – the 30 highest-activating images from the whole dataset according to the model's predictions ($N = 30$); (blue) suppress images – the 30 most suppressing/deactivating images according to the model's predictions ($N = 30$). Error bars show standard error of the mean across all images in each category. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. 29
2.9 **Iterative model training across days improves model prediction.** . . 30
2.10 **Prediction performance scales log-linearly with training data size.** Test R^2 as a function of training dataset size (number of unique images on x-axis) and number of repeats per image (different colored lines). Left: Representative targets showing individual scaling behavior. Right: Mean across $N = 23$ stable and reliable targets, normalized by noise ceiling. Shaded regions represent standard error of the mean across 10 random seeds (for individual targets on left) or across targets (for population average on right). Results are shown for adversarially trained ResNet-50. Test performance was evaluated on a fixed, separate day with 4 repeats per image, regardless of training setup. The relationship between test R^2 and number of training images is approximately linear on the log scale. Additionally, increasing the number of repeats per image in the training set improves downstream test performance. 31

2.11 **More unique images outperform more repeats for a fixed total budget.** Test R^2 as a function of total budget size in images (x-axis) and number of repeats per image (different colored lines). Identical to Figure 2.10, but with the x-axis showing the total budget size in images instead of the number of unique images. 32
2.12 **Neural sites show distinct patterns of semantic category selectivity.** For each representative neural target, images from the ImageNet validation set and TrailerFacesHQ dataset ($N = 50{,}000$ total images) were passed through the trained model to obtain predictions. Images were then grouped by their class labels, and the mean ± standard deviation of model predictions was computed for each class (shown as error bars on x-axis). We show the top- and bottom-10 image classes in terms of their mean predicted responses for each target. 33
2.13 **Example drive and suppress stimuli for representative neural sites.** Each panel (a–d) shows results for one neural site, with four rows of 10 example images each. Row 1 ("drive"): Images predicted by the model to maximally drive neural activity at the target site. Row 2 ("high"): Images from the first recording session (used for model training) that empirically elicited the highest neural responses. Row 3 ("low"): Images from the first recording session that empirically elicited the lowest neural responses. Row 4 ("suppress"): Images predicted by the model to maximally suppress neural activity. 34
2.14 **Multi-target optimization achieves selective neural modulation.** For two example neural sites, this figure shows target and off-target effects of noninvasive modulation, in the same format as Figure 2.8. Top row: optimization performed on a single target site without constraining off-target effects. Bottom row: optimization performed with multi-target constraints, driving or suppressing the target while keeping off-target responses within 0.3 standard deviations of the random image baseline. 36

3.1 **Hardware setup for neural recording and stimulus presentation.** A task laptop presents stimuli on a monitor via HDMI and logs behavioral responses via a USB mouse. A physics lab-grade photodiode (DET36A2) is placed on the monitor to capture precise stimulus onset timing; its analog signal is sent via BNC cable to the Blackrock NeuroPort for synchronization. The subject's implanted SEEG clinical leads connect to a custom splitter box, which passively splits the signals. One path goes to the clinical EEG system for standard hospital monitoring; the other path goes via touch-proof cables to a Blackrock amplifier, then over fiber optic to the NeuroPort, and finally via Ethernet to a central host PC for research recording. This configuration enables identical neural signals to be recorded in parallel by both the clinical system and the research system, time-locked to stimulus onset via the photodiode. . 38

3.2 **Detailed task paradigm showing the rapid serial visual presentation (RSVP) setup with the auxiliary attention task (dog detection).** Images from multiple datasets are presented sequentially at variable intervals. Patients maintain fixation on the central cross and respond with a mouse click whenever they detect a dog in the images. Correct detections within 2 seconds produce an upward sound; incorrect clicks (no dog in previous 2 seconds) produce a downward sound. At the end of each session, patients receive performance feedback and an in-game medal based on their dog detection accuracy. 40
3.3 **Automatic target extraction pipeline for patient MIT001.** Each panel corresponds to one neural site (bipolar electrode pair). The heatmap shows the split-half correlation coefficient R computed on the training split of images as a function of response window parameters: the x-axis shows the start time of the averaging window relative to image onset, and the y-axis shows the window duration. The star marks the parameter combination yielding the highest training split-half correlation. The text adjacent to each star indicates the test split-half correlation at those optimal parameters, providing an unbiased estimate of visual reliability for that target configuration. 44
3.4 **Automatic target extraction pipeline for patient MIT002.** Same format as Figure 3.3. 45
3.5 **Automatic target extraction pipeline for patient MIT003.** Same format as Figure 3.3. 46
3.6 **Visual reliability across all neural sites for patient MIT001.** Each panel shows one neural site (bipolar electrode pair). The x-axis shows time relative to image onset (0 ms). The y-axis shows the split-half correlation coefficient R. Red shading indicates time points where $p < 0.001$. This figure extends Figure 2.2 from the main text by showing all recorded electrodes rather than representative examples. 52
3.7 **Visual reliability across all neural sites for patient MIT002.** Same format as Figure 3.6. 53
3.8 **Visual reliability across all neural sites for patient MIT003.** Same format as Figure 3.6. 54

A.1 **Intra-day visual reliability remains stable across recording days.** Each bar represents the mean split-half correlation coefficient computed within a single recording day, averaged across all time steps (-0.1 to 0.9 s relative to image onset) and across all neural sites (without filtering for reliability). Error bars indicate standard deviation across neural sites. (a) Patient MIT002. (b) Patient MIT003. Neither patient showed statistically significant changes in intra-day reliability across recording days (all pairwise comparisons $p > 0.08$), suggesting that the capacity for visual encoding remains stable throughout the hospital stay. 60

A.2 **Raw voltage sample distributions across recording days for patient MIT002.** Each panel corresponds to one neural site (bipolar electrode pair). Histograms show the distribution of all raw voltage samples collected from each electrode on each recording day, with different colors representing different days. Distributions were Windsorized by removing the top and bottom 1% of samples to reduce the influence of outliers. Some electrodes exhibit remarkably stable distributions across days (nearly overlapping histograms), while others show changes in distribution width or shape. Electrodes that were used as drive/suppress targets in the main analyses are highlighted with red borders. 61
A.3 **Raw voltage distributions for patient MIT003.** Same format as Figure A.2. 62
A.4 **Relationship between distribution stability and cross-day reliability.** Each point: one electrode for one day pair. X-axis: cross-day stability for images (cross-day correlation for repeated images on different days divided by the split-half correlation on the first recording day). Y-axis: raw acquired samples distribution stability for that target ($\min(SD_a/SD_b, SD_b/SD_a)$). Both patients show significant positive correlation: electrodes with stable signal properties exhibit higher cross-day visual reliability. 63

B.1 **Raw photodetector output (y-axis, millivolts) as a function of time (x-axis).** This example displays the photodetector response to five successive image onsets. For each onset, the voltage rises sharply and then decays exponentially. 66
B.2 **Z-scored derivative of the photodetector signal for a representative time segment in an example recording session.** The y-axis shows the z-score of the temporal derivative; the x-axis shows time. Arrows correspond to ongoing image onsets. 66
B.3 **Event-triggered average of the photodetector signal, time-locked to high-threshold crossings.** The y-axis shows signal magnitude in millivolts; the x-axis shows time in seconds relative to the threshold crossing. The solid line indicates the mean across all crossings, and the shaded region indicates the standard deviation. 67
B.4 **Cross-correlation between the template waveform and the full recording session.** The y-axis shows the correlation coefficient; the x-axis shows time. Correlation values approach 1 at times when the recorded signal closely matches the template, corresponding to image onset times. 67
B.5 **Histogram of timing differences between expected task triggers from the task computer and photodetector-extracted triggers.** The y-axis shows the number of triggers; the x-axis shows offset in milliseconds. 68
B.6 **Validation of trigger-to-image correspondence.** The figure shows correlations between inter-trigger intervals from the photodetector pipeline and expected inter-stimulus intervals from the task computer. The analysis is performed across 20 time bins (columns), each representing 5% of the recording session. Five rows show correlations at different trigger offsets: -2, -1, 0, $+1$, and $+2$ (blue for negative offsets, black for zero, red for positive). 69

C.1 **Off-target effects of drive and suppress stimuli across the brain.** Each row corresponds to a different neural site being targeted for driving or suppression. Points represent neural sites in an orthographic brain map projection. Color indicates the mean activity of each site in response to the drive/suppress stimuli designed for the target site, normalized by that site's response to random images (z-scored relative to mean and standard deviation). The large dot (2× size) indicates the target site for which stimuli were optimized. All other points show off-target effects: how stimuli optimized for one site affect activity at other sites across the brain. 72

List of Tables

3.1 Recording session parameters for each patient. The "Day" column refers to post-operative day. 39

Chapter 1

Introduction

The ability to collect large-scale datasets of brain responses to corpuses of naturalistic sensory stimuli has opened up new possibilities in systems neuroscience [1]. In particular, this advance has enabled the creation of computational models that predict neural responses in a purely data-driven manner. Deep neural network models trained to predict neural activity have proven remarkably effective at capturing response properties of visual cortex [2–5] and higher-level cognitive brain regions such as the language network [6–10]. These trained models can be effectively used to extract scientific knowledge about the processing of information in the brain [11–13], and provide insights into feature selectivity and representational geometry that were previously inaccessible. Moreover, the models have been leveraged as tools for neural control: they allow researchers to design optimal stimuli that drive or suppress targeted neural populations. This approach has been pioneered in non-human primates to control neural activity in inferotemporal (IT) cortex [14–16].

Here, we extend this model-based neural control framework to the human visual system using stereotactic electroencephalography (SEEG). SEEG, recorded from patients undergoing clinical monitoring for epilepsy, combines the high temporal resolution of electrophysiology with the spatial specificity of intracranial recordings in the human brain.

The ability to selectively control neural activity in the human brain may have profound implications for next-generation brain-computer interfaces and treatments of neurological disorders. A non-invasive approach that could drive or suppress targeted neural activity patterns through carefully designed sensory stimuli would open new therapeutic avenues for epilepsy patients as well as for those with conditions ranging from chronic pain and PTSD to depression and anxiety, where maladaptive neural patterns may be selectively modulated. Beyond clinical applications, spatially and temporally precise neural control would enable causal testing of computational theories of human cognition, allowing researchers to perturb neural populations of interest. This work establishes that such control is achievable in the human brain using model-based visual stimulus optimization. By demonstrating that deep learning models trained on intracranial recordings can successfully predict which natural images will drive or suppress activity at specific cortical sites, we provide a proof-of-concept for stimulus-based neural control in humans. While the current implementation relies on SEEG recordings in surgical patients, the framework generalizes to other modalities of neural recording (EEG, fMRI, fUSI, etc.) and other sensory modalities (audio, video, etc.).

Our approach proceeds in three stages, following prior work [13,14]. First, we collect

SEEG responses to a large corpus of natural images, establishing a dataset of visually-evoked neural activity at many simultaneously recorded sites across the brain. Second, we train encoding models that predict SEEG responses based on deep neural network representations of images. Then, we use these models to identify new images that are predicted to selectively modulate activity at targeted recording sites, and finally we test these predictions in our experimental setup.

Contributions

In this work, we make four core contributions:

1. We establish a map of visual selectivity across regions in the human brain, revealing which areas have reliably predictable visual responses.

2. We introduce the first dataset of intracranial EEG (iEEG) responses to rapid serial visual presentation (RSVP) of natural images in humans.

3. We demonstrate that our model-guided framework can successfully identify visual stimuli that selectively drive and suppress activity at specific neural sites. Moreover, we show that targeted stimulation can modulate one site without strongly affecting other sites, establishing the selectivity of neural control.

4. We systematically characterize the scaling laws of model performance as a function of both dataset size and the number of image repeats, providing principled guidance for future experimental design.

These results establish the feasibility of model-based visual stimulus optimization for human intracranial electrophysiology, opening new avenues for investigating the structure of human visual representations.

1.1 Related Work

Predicting brain responses to sensory stimuli using pretrained models

Using pretrained models to predict brain responses to sensory stimuli is a growing field of research, with various pretrained vision models (such as ResNet-50, CLIP, DINO, ViT, etc.) being used to predict brain responses to sensory stimuli [2–5,17].

For language, large language models (LLMs) have been used to predict responses in the human language network and other fMRI targets to language stimuli [8,9]. While fMRI allows for building models of the responses in the human brain, the temporal resolution of fMRI is limited to seconds. This limitation is addressed in our work with SEEG, which offers millisecond temporal resolution and the spatial specificity of individual electrode contacts.

Model-based neural control

Our work builds on a growing body of research using computational models for neural control. In non-human primates, deep neural network models trained on neural responses to natural

images have enabled the synthesis of images that precisely modulate activity in specific populations of IT cortex neurons [12,14–16]. There exists similar work in mice [13]. Our work extends this framework to the human brain.

Recent work has extended the control framework to human neural responses to language using functional magnetic resonance imaging (fMRI; [6,7]). Our work extends this framework to visual stimuli in the human brain using SEEG.

Capturing neural responses to visual stimuli using SEEG

While there are plenty of existing human datasets of fMRI responses to language [18–20] and visual [21,22] stimuli, including multimodal datasets which can combine video and audio [23–25] and also similar EEG datasets [26–28], datasets using human intracranial electrophysiology (SEEG, ECoG, etc.) are very sparse. There exists a large dataset in ECoG for pose estimation (AJILE; [29,30]) and smaller SEEG datasets of human language production [31,32] and multimodal stimuli (movies; [33]). The closest existing dataset for our purposes is the BrainTreebank dataset [34], which was collected in a hospital environment using naturalistic stimuli (movies). However, it did not have the necessary temporal precision of the alignment of neural data to stimulus presentation (see Appendix B for our trigger extraction pipeline).

Scaling laws of neural prediction

The scaling laws analysis, i.e. predicting the performance of the model in relation to dataset and model size, is a well-established technique in the field of nature language processing [35,36]. More recently, this technique has been applied to models predicting neural responses to sensory stimuli [10,37] and also in the self-supervised setting when predicting masked regions of neural activity [37–40]. Our work investigates the scaling laws of neural prediction performance as a function of SEEG dataset size (see Section 2.6).

Chapter 2

Results

In this section, we present the results of our study. We investigated the visual reliability of neural responses across brain regions (Section 2.1), the stability of the responses over time (Section 2.2), the performance of our predictive models (Section 2.3), and our model-guided neural control framework (Section 2.4). Furthermore, we studied the iterative model improvement (Section 2.5), scaling laws of prediction performance (Section 2.6), semantic categories of drive and suppress stimuli (Section 2.7), and finally we investigated the spatial selectivity of our stimulation protocol across multiple targets (Section 2.8).

2.1 Visual reliability varies across brain regions

To investigate which brain regions can be reliably modulated by visual stimuli, we recorded stereotactic electroencephalography (SEEG) signals from epilepsy patients viewing natural images in a rapid serial visual presentation (RSVP) experiment setup (Figure 2.1a). The experimental setup consisted of subjects viewing successive images on a screen in front of them while neural data was recorded using SEEG probes. We used images from multiple datasets, such as OASIS [41], ImageNet [42], and TrailerFacesHQ [43]. These images were presented rapidly in succession with variable time intervals between them, such that on average, each image was presented every 250 milliseconds (Figure 2.1b). To monitor subjects' attention throughout the data acquisition sessions and increase subject engagement, we included an auxiliary task (described in detail in Methods). Throughout the experiment, subjects were instructed to maintain fixation on a cross presented at the center of the screen. Figure 2.1c shows a diagram of our model-based experimental approach.

In each subject, neural activity was recorded across around 100 SEEG electrode contacts from $n = 469$ to $n = 1319$ images, with each image repeated for a total of $m = 4$ to $m = 12$ times within a single recording session. Each pair of adjacent SEEG electrode contacts defines a neural site (we use the bipolar montage, re-referencing the signal by taking the difference between adjacent electrode contacts; see Methods, Section 3). We quantified the visual reliability of each neural site using a split-half correlation approach (Figure 2.2a). For a given neural site, suppose each of N images is presented m times. For each image, we randomly divide its m repeats into two halves. For each half, we average the neural responses for every image, producing two N-dimensional vectors, $\mathbf{V}(t)$ and $\mathbf{V}'(t)$. The responses are

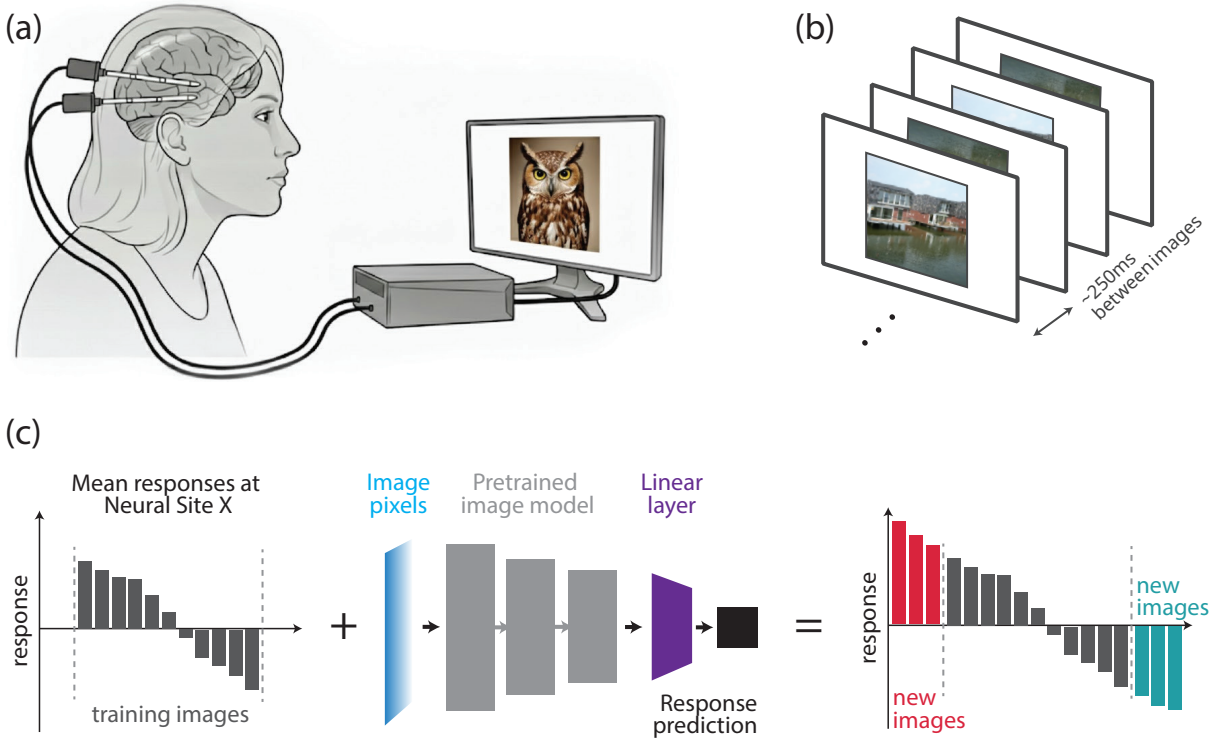


Figure 2.1: **Experimental setup, model, and key questions.** (a) Schematic of a subject with implanted SEEG probes viewing successive images on the screen. Neural data was recorded as subjects viewed images in a rapid serial visual presentation (RSVP) setup. (b) The RSVP task. Images from various datasets (OASIS, ImageNet, and TrailerFacesHQ) were presented in rapid succession (on average every 250 ms), with subjects required to maintain fixation on a central cross and perform an auxiliary attention task (see Methods). (c) Overview of our model-based experimental approach and main questions. On the left, a sketch shows a set of responses to training images being provided as input to a pre-trained neural network model (such as a convolutional neural network, depicted as layers), followed by a linear readout layer that is trained to predict the neural response recorded at a specific SEEG site. Once trained, this model enables us to take a large set of candidate images and predict which stimuli will most strongly drive or suppress the targeted neural site.

measured as a function of time relative to image onset (which corresponds to timepoint $t = 0$). The split-half correlation is then defined as the Pearson correlation coefficient between these two vectors across all N images. The underlying assumption is that if the same image is presented twice, the neural response of a reliable neural site should be the same. A high split-half correlation would indicate that most of the response variance at that neural site is explained by variability in the visual input, whereas a correlation near zero would suggest that neural activity mostly varies independently of the visual input.

Figure 2.2b shows the visual reliability of four representative neural sites. We found that visual reliability emerged shortly after image onset and persisted throughout the presentation period, with reliability persisting well beyond the image offset, all the way up to about 1 second after image onset. The reliability increased predictably with the number of image repeats that were averaged together (Figure 2.2b, right panels). Shuffling the image order between the two halves abolished these correlations (gray dashed lines in Figure 2.2b).

To identify the optimal response time windows for each neural site, we used an automated target extraction pipeline that sweeps over different window parameters (see Methods, Section 3.7).

Across the brain, visual reliability was highest in occipital and temporal regions (Figure 2.3). This distribution is consistent with the known hierarchy of visual processing, with early visual areas in occipital cortex and higher-level areas in temporal cortex showing the strongest stimulus-driven responses [44,45].

2.2 Stability of neural targets over time

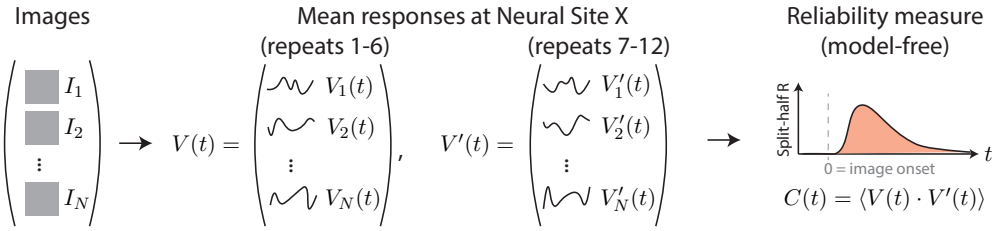
Having established that many neural sites show reliable visual responses within a recording session, we next asked whether these representations remain stable across multiple days.

We compared neural responses to $n = 55$ images which were repeatedly presented on three different recording days for each subject. In Figure 2.4, the scatter plots show the neural response to each image on the first recording day (Day 2; y-axis), compared against: (1) the response to the same image presented again on the same day (left panel, x-axis), (2) the response on Day 5 (middle panel, x-axis), and (3) the response on Day 7 for subject MIT003 or Day 13 for subject MIT002 (right panel, x-axis). Each dot represents one image that was shown to the subject on the different recording days. In this section, "neural response" means the averaged response across 4 repeats per image.

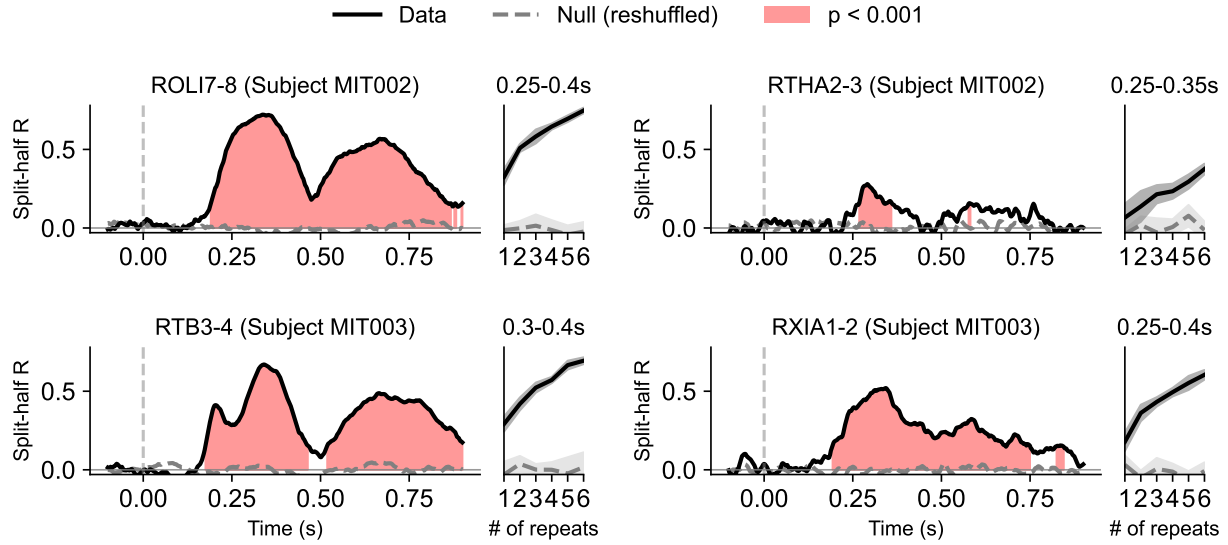
Some neural sites, such as ROLI7-8, maintained remarkably stable representations throughout the subject's hospital stay, with correlations remaining high across all recording sessions ($r = 0.75, 0.76, 0.78$ on Days 2, 5, and 13, respectively). Other sites, such as RTHA2-3, showed degraded stability on later days, with correlations diminishing substantially over time ($r = 0.44, 0.18$ on Days 2 and 13, respectively).

We investigated what could be the reason for decreased stability of targets across days, and found that the stability of the targets is correlated with the stability of the standard deviation of the distribution of the raw acquired data samples from our equipment (Figure A.4b and A.4a).

We define *reliable targets* as those with split-half $R \geq 0.4$ at 4 repeats within a recording session, and *stable targets* as those with cross-day correlation $R \geq 0.3$ and where R is at least



(a) The split-half correlation method for computing visual reliability. For a selected neural site (electrode pair), images I_1 through I_N are presented for up to 12 repeats each. Half of the randomly selected repeats are averaged to obtain response vectors $V_1(t), V_2(t), \ldots, V_N(t)$, collectively forming an N-dimensional vector $\mathbf{V}(t)$. The same procedure is applied to the remaining half of the repeats to obtain $\mathbf{V}'(t)$. The split-half correlation $C(t)$ is the Pearson correlation coefficient between $\mathbf{V}(t)$ and $\mathbf{V}'(t)$.



(b) Reliability of visual responses in representative neural sites. Panels show data from different neural sites (SEEG electrode pairs) in two subjects. Left: Split-half correlation coefficient R as a function of time relative to image onset ($t = 0$). The colored regions indicate time points where $p < 0.001$, showing when visual reliability is statistically significant. Right: Split-half correlation as a function of the number of image repeats averaged together. For example, at 1 repeat, each image was presented once in each half; at 6 repeats, each image was presented 6 times per half and the responses were averaged. Gray shaded regions show standard deviation across randomizations of the repeat ordering. Gray dashed lines in both columns show the same results for null distributions obtained by shuffling the image order between the two halves.

Figure 2.2: **Split-half correlation reveals neural sites with reliable visual encoding in the human cortex.** For results in all neural sites, see Figures 3.3, 3.4, 3.5.

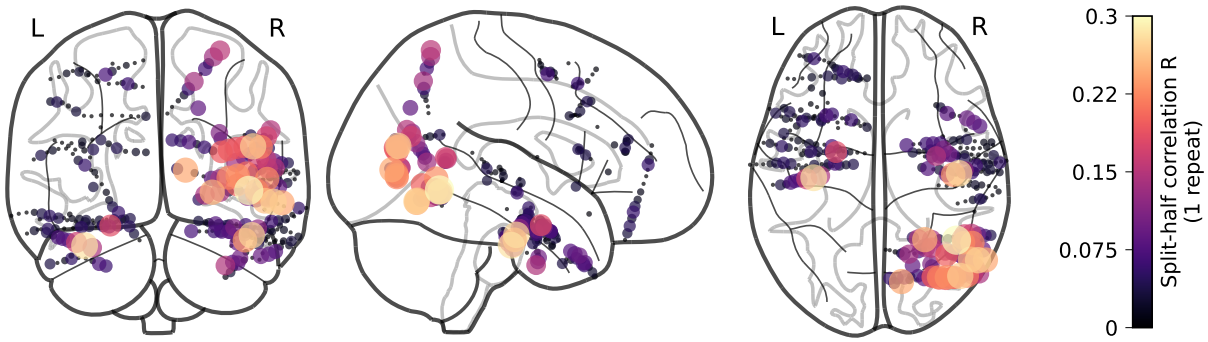


Figure 2.3: **Visual reliability is highest in occipital and temporal cortex.** Orthogonal brain map showing all SEEG neural sites as dots. Dot size and color represent the split-half correlation coefficient R computed using one repeat per image. Highest reliability is observed in occipital and temporal regions. Based on the measure from Figure 2.2.

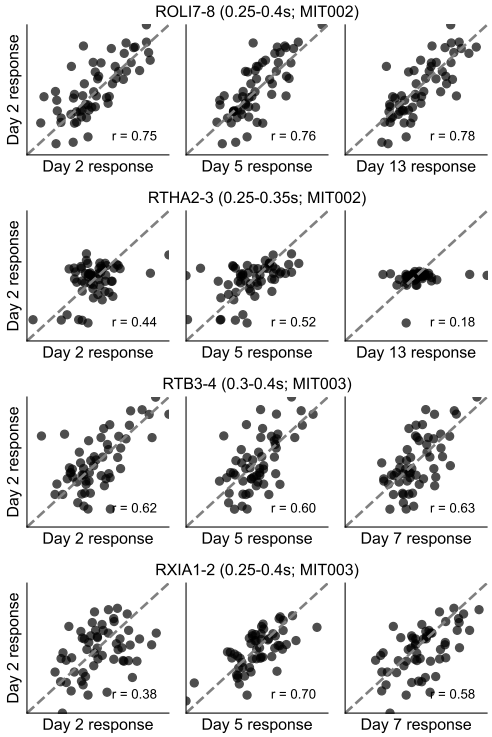


Figure 2.4: **Some neural sites maintain stable representations across days while others degrade.** Scatter plots for four example neural sites showing response correlations across days. For each target, the y-axis shows the neural response on the first recording day (Day 2). The x-axis shows: (left) the response to the same image presented again on the same day; (middle) the response on Day 5; (right) the response on Day 7 (subject MIT003) or Day 13 (subject MIT002). Each dot represents one image. Some targets like ROLI7-8 maintain high correlations across all days, demonstrating remarkable stability throughout the subject's hospital stay. Other targets like RTHA2-3 show substantially degraded stability over time. In this figure, each response was averaged across 4 repeats per image.

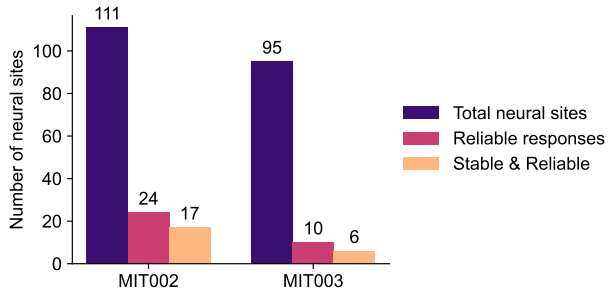


Figure 2.5: **A minority of neural sites are both reliable and stable across days.** Bar plot showing the proportion of neural sites in each experimental subject across three categories: total sites (all electrode pairs), reliable sites, and stable+reliable sites (see Main Text and Methods for definitions). Only a subset of all sites show reliable visual responses, and only a further subset of those maintain stability across recording days.

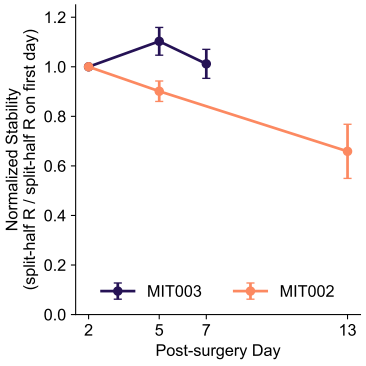


Figure 2.6: **Neural target stability varies over time post-surgery.** Y-axis shows stability, defined as the correlation of neural responses for the same set of images presented on different recording days, divided by the split-half correlation on the first recording day (at 4 repeats per image). All stability values are further normalized by dividing by their value on the first recording day. X-axis shows the post-surgery day. Only reliable targets are included in this analysis.

0.75× the within-day correlation on the first recording day. Notably, only approximately 10% of all neural sites were both reliable and stable across recording days (Figure 2.5).

Within the population of $N = 23$ stable and reliable targets across two subjects, stability varied over time post-surgery. In Figure 2.6, the y-axis shows stability (defined as the cross-day correlation normalized by the split-half correlation on the first recording day), normalized to the first recording day. For the subject MIT002, there was a clear trend of stability decreasing over time post-surgery, while for MIT003, there was no clear trend.

In the rest of this study, we focus on those $N = 23$ stable and reliable sites, and report combined population results for all of the sites selected at this stage.

2.3 Predicting neural responses from image features

Beyond simply measuring visual response reliability, we next asked whether we could predict the neural response as a function of the image pixels (Figure 2.7). We used pretrained deep neural network models (Adversarially Trained ResNet-50, ResNet-50, CLIP ViT-B/32, DINO ViT-S/16, and ViT-B/16; described in Methods Section 3.9) to extract image features, and trained linear regression models to predict neural responses at each site as a function of the extracted features.

Test performance was evaluated using sessions recorded on future days after the training session, providing a strong test of model generalization. Different pretrained models varied in their ability to predict neural responses (Figure 2.7, right). The highest predictive performance was achieved by the adversarially trained ResNet-50, with an average test $R^2 = 0.41$, consistent with prior work in non-human primate modeling of the visual stream [15,16].

2.4 Driving and suppressing neural activity with model-selected stimuli

Given that models can predict neural responses to arbitrary images, we tested whether we could use these predictions to select stimuli that drive or suppress activity at specific neural sites.

For each target, we used our trained models to predict responses to all images in our bank of images ($n > 50000$ images). We then selected the 30 images with the highest predicted responses ("drive" stimuli) and the 30 images with the lowest predicted responses ("suppress" stimuli). These model-selected images were then presented to subjects in subsequent recording sessions to validate whether the model predictions matched the actual neural responses.

In the follow-up recording sessions, our model-selected stimuli successfully modulated neural activity in predicted directions (Figure 2.8). Drive stimuli elicited significantly higher responses than random images, while suppress stimuli elicited significantly lower responses. Due to limited availability of our study participants for research sessions, we show results for 6 targets, which were all of the stable and reliable targets that we tried to drive and/or suppress. The candidate drive/suppress targets were selected independently beforehand, without the ability to see the drive and suppress results.

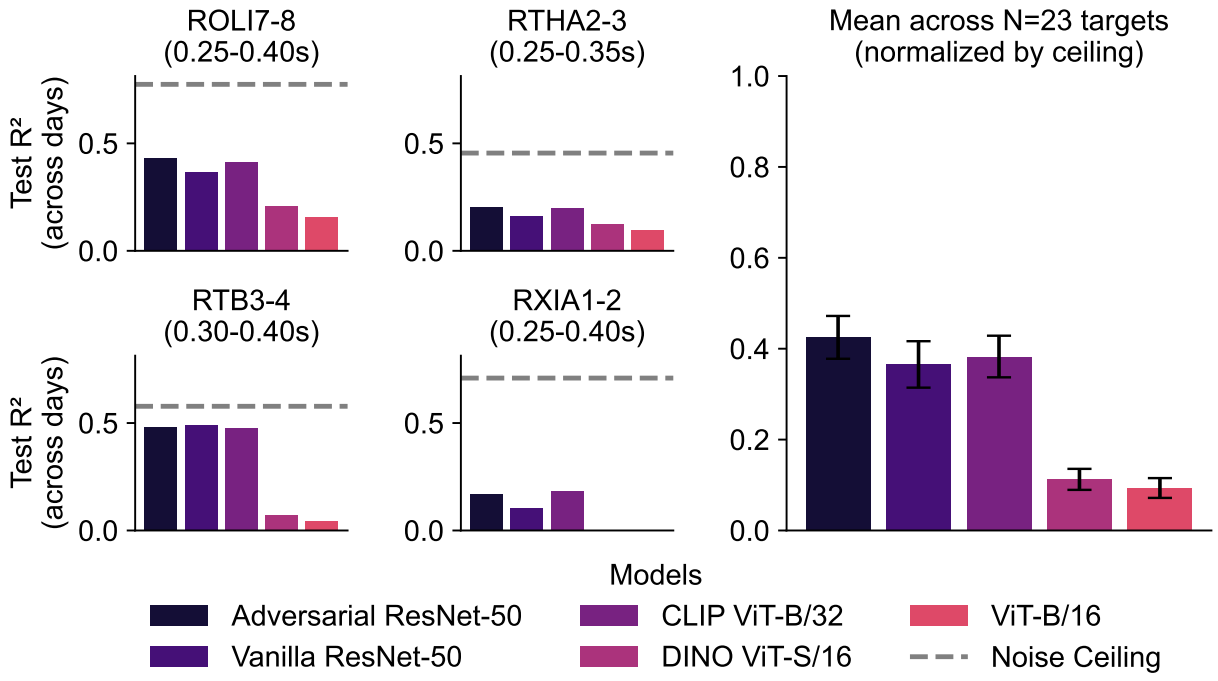


Figure 2.7: **Adversarially trained ResNet-50 achieves highest prediction of neural responses.** Left: Test R^2 for four example neural sites across different pretrained architectures. Right: Mean test R^2 across $N = 23$ stable and reliable targets, normalized by the noise ceiling of each target. Each bar corresponds to a different pretrained image model. Training and test sessions were conducted on different days, providing a strong test of model generalization and ensuring that models capture stable aspects of neural representations. All test data consisted of 4 repeats per image, regardless of the training setup. Error bars represent standard error of the mean across targets.

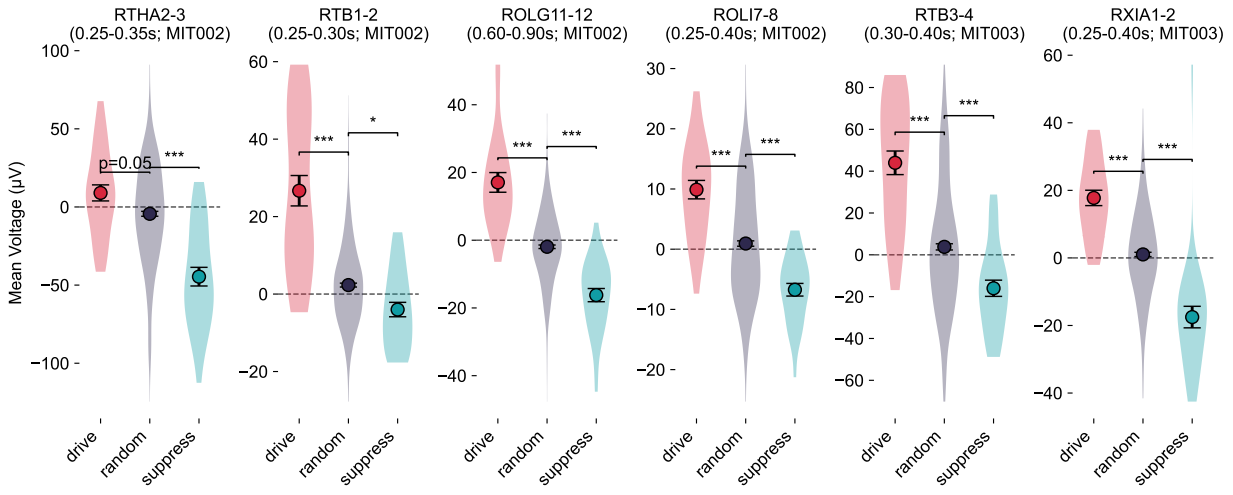
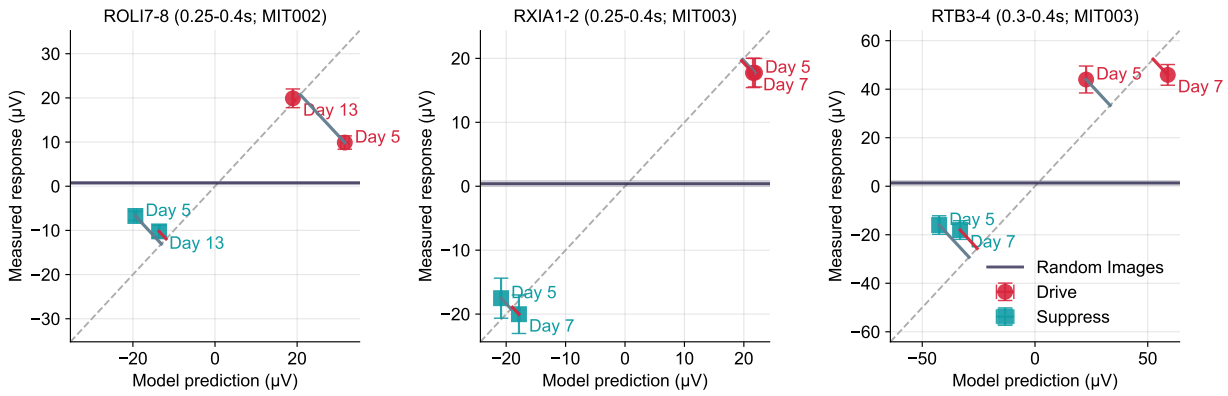


Figure 2.8: **Model-selected stimuli successfully drive and suppress neural activity.** Each panel shows one neural target. Three bars represent mean voltage in the target neural site for different image categories: (gray) random images – randomly selected new images shown on that day ($N > 200$); (red) drive images – the 30 highest-activating images from the whole dataset according to the model's predictions ($N = 30$); (blue) suppress images – the 30 most suppressing/deactivating images according to the model's predictions ($N = 30$). Error bars show standard error of the mean across all images in each category. Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

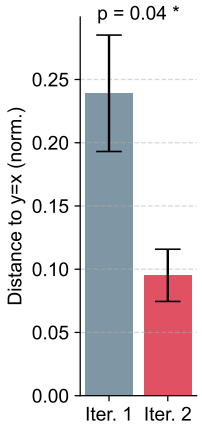
2.5 Iterative model improvement with feedback data

We hypothesized that our predictive models could be improved by including additional training sessions, which contain neural responses to the drive and suppress stimuli. In other words, we hypothesized that by observing how the brain actually responds to the images that the model predicted would drive or suppress activity, the model can be refined in subsequent iterations to produce more accurate predictions for new images. To test this hypothesis, we compared models used in early sessions (Iteration 1, tested on day 5 post-op) to models retrained from scratch on all available data from all previous sessions (Iteration 2, tested on days 7 and 13 post-op).

For the targets that remained both reliable and stable across all recording sessions and contained drive/suppress data (only 3 targets met these criteria), models from later iterations produced predictions that were closer to the actual neural responses (Figure 2.9a). We quantified this effect by computing the perpendicular distance from data points to the $y = x$ line (which represents perfect prediction), and found that when pooling data from all three targets, Iteration 2 models showed significantly lower distances (better predictions) than Iteration 1 models ($p < 0.05$; Figure 2.9b). We normalized every distance by the maximum prediction or response on that day to enable pooling across targets. This improvement suggests that the additional data collected from previous iterations can be included in the training set to improve the model's performance in the future iterations.



(a) Model predictions vs. measured responses across iterations. Comparing model-predicted responses (x-axis) to actual measured neural responses (y-axis) for three targets that remained both reliable and stable across all recording sessions and contained drive/suppress data. Each data point represents the mean response to $N = 30$ drive or suppress images; error bars show standard error of the mean. Black horizontal lines show mean ± SEM for $N > 200$ random images presented on different days. The dashed diagonal line indicates perfect prediction ($y = x$). Data points from later iterations (shown in different colors/symbols for different recording days) hug the $y = x$ line more tightly than earlier iterations, suggesting that models improve over time by incorporating feedback data from previous sessions.



(b) Quantification of prediction accuracy improvement across iterations. Y-axis shows normalized perpendicular distance from data points to the $y = x$ line (lower values indicate better predictions closer to perfect agreement). Distance was normalized by dividing by the maximum model prediction or measured response on each day, allowing results to be pooled across targets. Bars represent the mean and standard error of the mean of the perpendicular distance across all image conditions (drive and suppress) and targets.

Figure 2.9: **Iterative model training across days improves model prediction.**

2.6 Scaling laws of neural prediction

To investigate how prediction performance scales with data quantity, we systematically varied both the number of unique images and the number of repeats per image in the training set. By doing so, we obtained the scaling laws of prediction performance in data quantity (Figure 2.10).

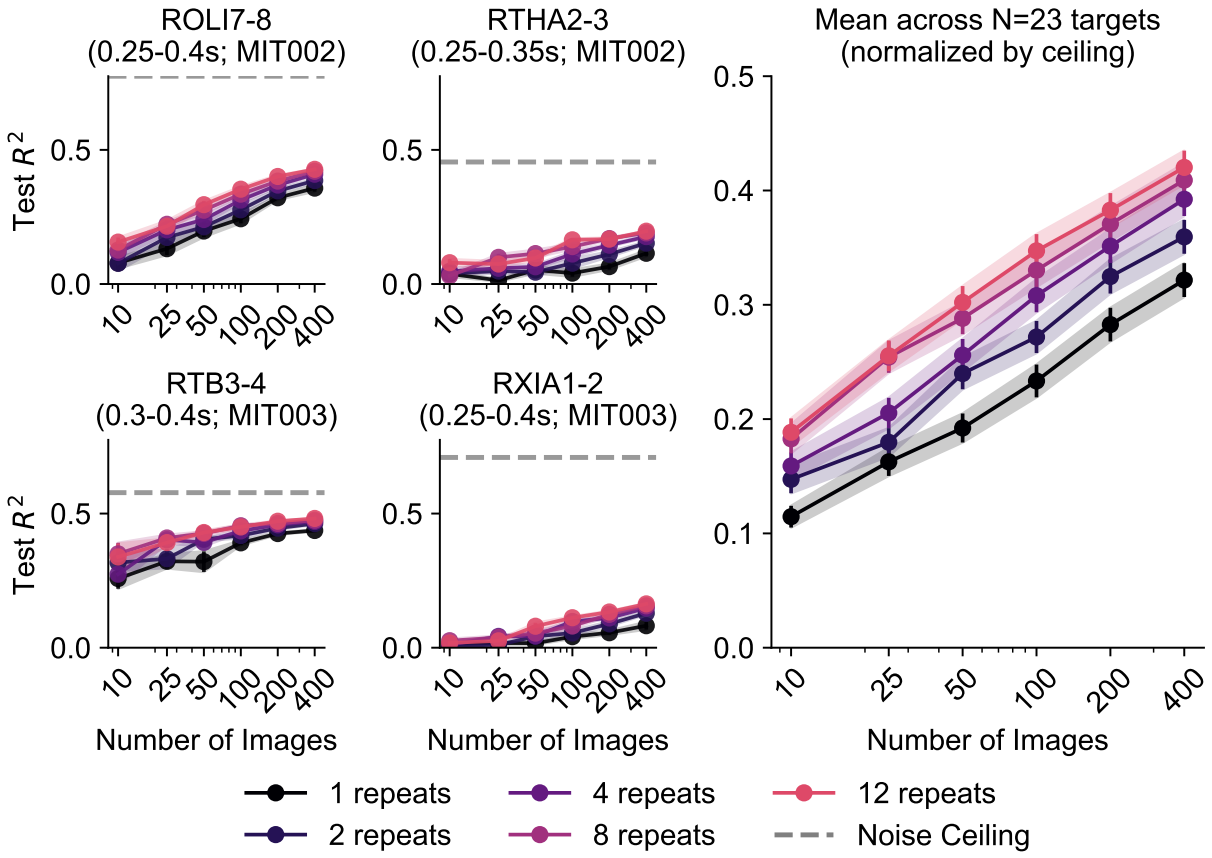


Figure 2.10: **Prediction performance scales log-linearly with training data size.** Test R^2 as a function of training dataset size (number of unique images on x-axis) and number of repeats per image (different colored lines). Left: Representative targets showing individual scaling behavior. Right: Mean across $N = 23$ stable and reliable targets, normalized by noise ceiling. Shaded regions represent standard error of the mean across 10 random seeds (for individual targets on left) or across targets (for population average on right). Results are shown for adversarially trained ResNet-50. Test performance was evaluated on a fixed, separate day with 4 repeats per image, regardless of training setup. The relationship between test R^2 and number of training images is approximately linear on the log scale. Additionally, increasing the number of repeats per image in the training set improves downstream test performance.

We observed an approximately linear relationship between test R^2 and the number of training images on a log scale (Figure 2.10). Additionally, increasing the number of repeats

per image in the training set improved downstream test performance substantially, suggesting that reducing noise in the training set by averaging over multiple presentations yields more accurate neural response estimates for model fitting. These scaling laws provide guidance for experimental design: both the diversity of stimuli (number of unique images) and the reliability of measurements (repeats per image) contribute to model performance. We also examined the predictive performance as a function of the total image budget size (Figure 2.11). The total budget size is the product of the number of unique images and the number of repeats per image. The relationship between test R^2 and the total image budget size was also approximately linear on the log scale. We noted, however, that for a fixed total image budget, increasing the number of unique images yields higher test R^2 than increasing the number of repeats per image, suggesting that the number of unique images is a more important factor for model performance than the number of repeats per image.

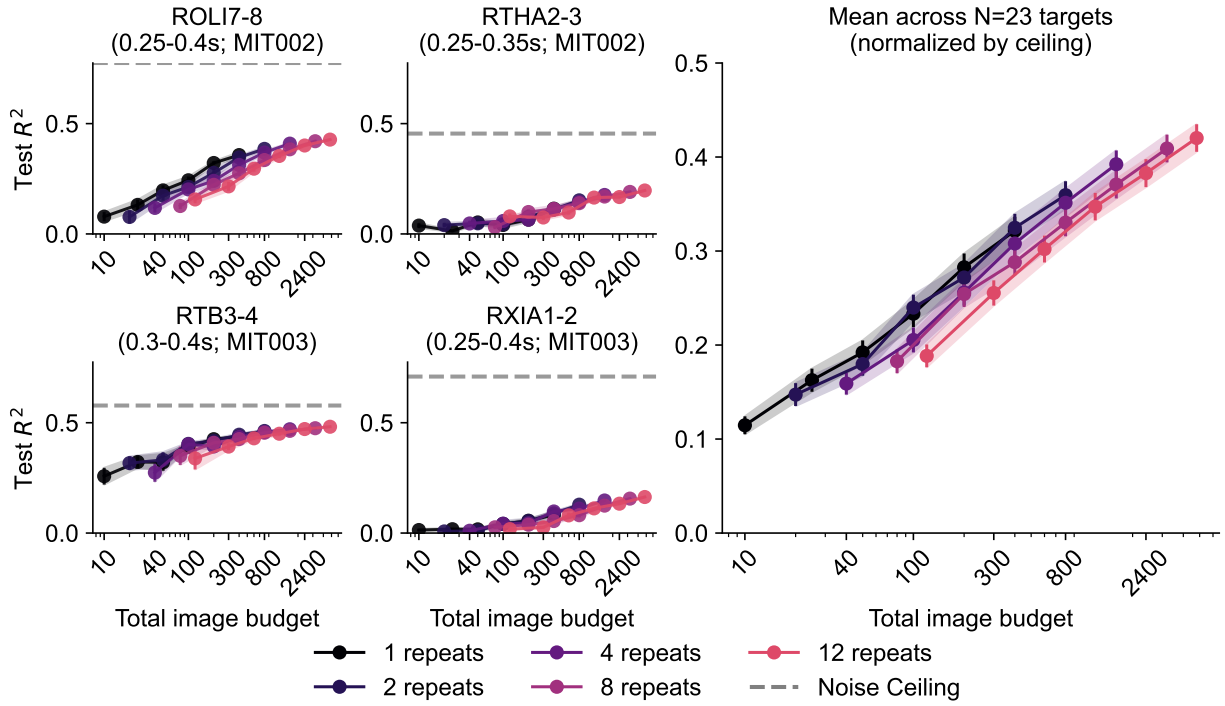


Figure 2.11: **More unique images outperform more repeats for a fixed total budget.** Test R^2 as a function of total budget size in images (x-axis) and number of repeats per image (different colored lines). Identical to Figure 2.10, but with the x-axis showing the total budget size in images instead of the number of unique images.

2.7 Interpreting model selectivity

We next sought to interpret the visual features that the models learned to associate with each neural site. We investigated which semantic categories of images tended to drive or suppress each target by passing all of the images from the ImageNet validation set and the TrailerFacesHQ dataset ($N = 50{,}000$ images total) through each of our trained models.

We then grouped the images according to their class labels and computed the mean and standard deviation of the models' predicted responses for each class. This analysis provides interpretable summaries of the selectivity patterns captured by the regression models.

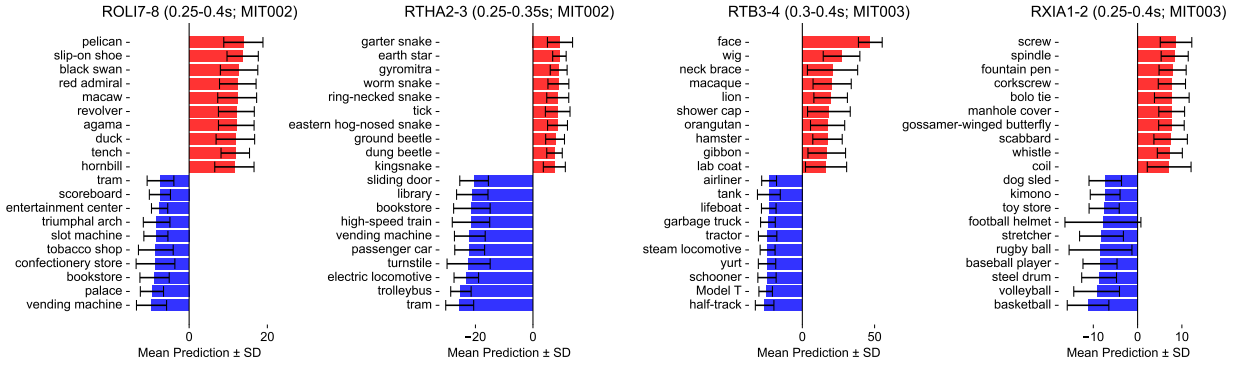


Figure 2.12: **Neural sites show distinct patterns of semantic category selectivity.** For each representative neural target, images from the ImageNet validation set and TrailerFacesHQ dataset ($N = 50{,}000$ total images) were passed through the trained model to obtain predictions. Images were then grouped by their class labels, and the mean ± standard deviation of model predictions was computed for each class (shown as error bars on x-axis). We show the top- and bottom-10 image classes in terms of their mean predicted responses for each target.

Different neural sites showed distinct patterns of class selectivity (Figure 2.12). For each target, we identified the 10 image classes with the highest mean predicted responses (most driving) and the 10 classes with the lowest mean predicted responses (most suppressing). This analysis revealed the semantic categories and provided interpretable summaries of the potentially complex selectivity patterns captured by our predictive models.

To further explore the nature of model predictions, we visualized example drive and suppress stimuli for four representative neural sites (Figure 2.13a-d). For each site, we show the model-predicted drive and suppress images against the images that empirically elicited the highest and lowest neural responses in the first training session (the same images that were used to train the model). Some of the sets show a clear pattern of similarity between the different drive or suppress images. For example, for RTB3-4 from subject MIT003, there was a clear pattern where all of the drive stimuli contained faces (Figure 2.13c, top); likewise, we conjecture that as drive stimuli for RTHA2-3 in subject MIT002, the model selected insect-like looking stimuli (even though many of them were not actually insects, Figure 2.13b, top); all of the suppress images for the site ROLI7-8 from subject MIT002 have a clear repeating visual pattern (although the images come from different classes, Figure 2.13a, bottom). However, some sets of stimuli like the drive images for ROLI7-8 (Figure 2.13a, top) and RXIA1-2 (Figure 2.13d, top) did not have such clear predictable patterns.

2.8 Selectivity of the stimulation protocol

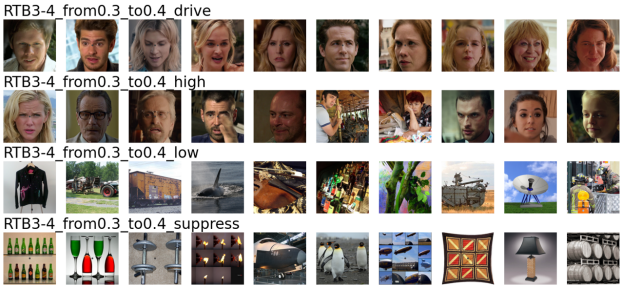
Having demonstrated that model-selected stimuli can drive and suppress individual neural sites, we next asked whether this modulation could be made selective: that is, whether we



(a) Neural site ROLI7-8 (subject MIT002).



(b) Neural site RTHA2-3 (subject MIT002).



(c) Neural site RTB3-4 (subject MIT003).



(d) Neural site RXIA1-2 (subject MIT003).

Figure 2.13: **Example drive and suppress stimuli for representative neural sites.** Each panel (a–d) shows results for one neural site, with four rows of 10 example images each. Row 1 ("drive"): Images predicted by the model to maximally drive neural activity at the target site. Row 2 ("high"): Images from the first recording session (used for model training) that empirically elicited the highest neural responses. Row 3 ("low"): Images from the first recording session that empirically elicited the lowest neural responses. Row 4 ("suppress"): Images predicted by the model to maximally suppress neural activity.

could modulate a target site while minimizing effects on other neural sites (off-target effects).

First, we examined the off-target effects for the targets from Figure 2.8. The results can be seen in Figure C.1. While most of the time, the target effects were stronger than the off-target effects, the off-target effects were not always negligible. This result is not surprising, as images were selected solely to maximize or minimize the predicted response at a single target site, without any regard to what happens at other sites.

Aiming to minimize these off-target effects, we developed an alternative optimization approach, where images were selected to drive or suppress the target site while simultaneously constraining off-target responses at another site to remain within 0.3 standard deviations of the random image baseline (Figure 2.14). These results are preliminary. In our exeriments, target sites showed robust driving and suppression effects while off-target sites remained near baseline levels. However, in the two sites that we experimented with, the original approach showed similar satisfactory results (Figure 2.14, top row) to the multi-target approach (Figure 2.14, bottom row).

While the end result was not significantly different from the single-target approach, our results suggest that the image selection framework can be extended to achieve spatially selective neural modulation, opening the possibility of independently controlling multiple neural populations through carefully chosen visual stimuli.

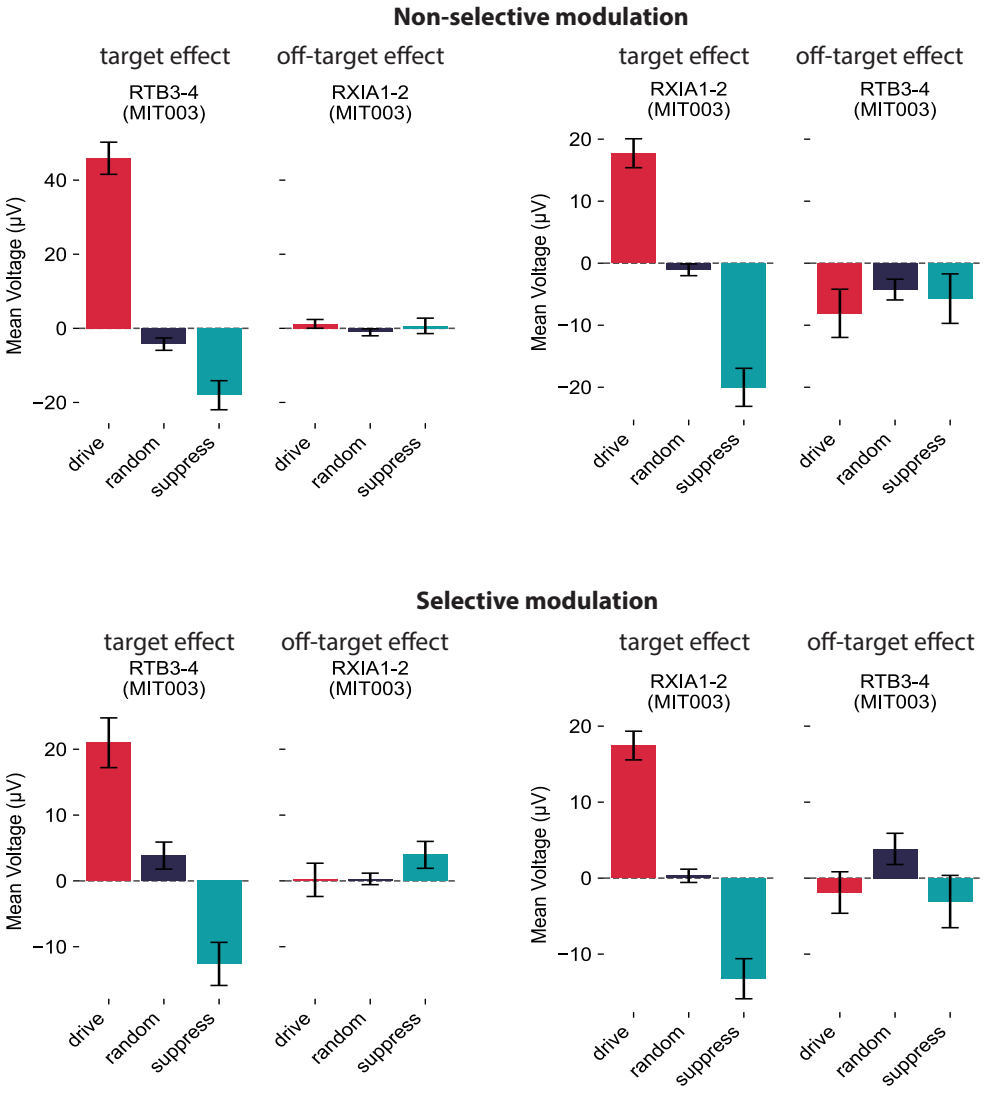


Figure 2.14: **Multi-target optimization achieves selective neural modulation.** For two example neural sites, this figure shows target and off-target effects of noninvasive modulation, in the same format as Figure 2.8. Top row: optimization performed on a single target site without constraining off-target effects. Bottom row: optimization performed with multi-target constraints, driving or suppressing the target while keeping off-target responses within 0.3 standard deviations of the random image baseline.

Chapter 3

Methods

3.1 Study Overview and Participants

Participant Cohort

We enrolled $N = 3$ patients undergoing stereotactic electroencephalography (SEEG) monitoring for medically refractory epilepsy at the Beth Israel Deaconess Medical Center (BIDMC). All patients provided informed consent under an IRB protocol approved at BIDMC.

Recording Timeline

Under the SEEG monitoring clinical protocol, patients remained in the hospital for up to 30 days for SEEG monitoring in the epilepsy monitoring unit. Research recording sessions began on post-operative day 2 and continued on multiple days until the patient was explanted and discharged. The number and duration of sessions depended on clinical constraints and patient availability. The earliest recording day for this study was post-operative day 2 and the latest was post-operative day 13 (patient MIT003). Recordings were conducted at various times of day depending on patient availability and clinical schedules.

3.2 Recording Setup

Stereo-EEG Probe Specifications

Each probe contained 6 to 16 contacts. The total number of channels ranged from 100 to 130 per patient across all probes. Electrode placement was determined solely by clinical needs; no additional electrodes were implanted for research purposes.

Recording System

Neural signals were recorded using a Blackrock NeuroPort Neural Signal Processor (256 channels) with a sampling rate of 10,000 Hz. The reference electrode was placed on the mastoid. We used a bipolar montage for all analyses, computed by taking the difference between adjacent electrode contacts. Each neural site corresponds to one bipolar electrode pair.

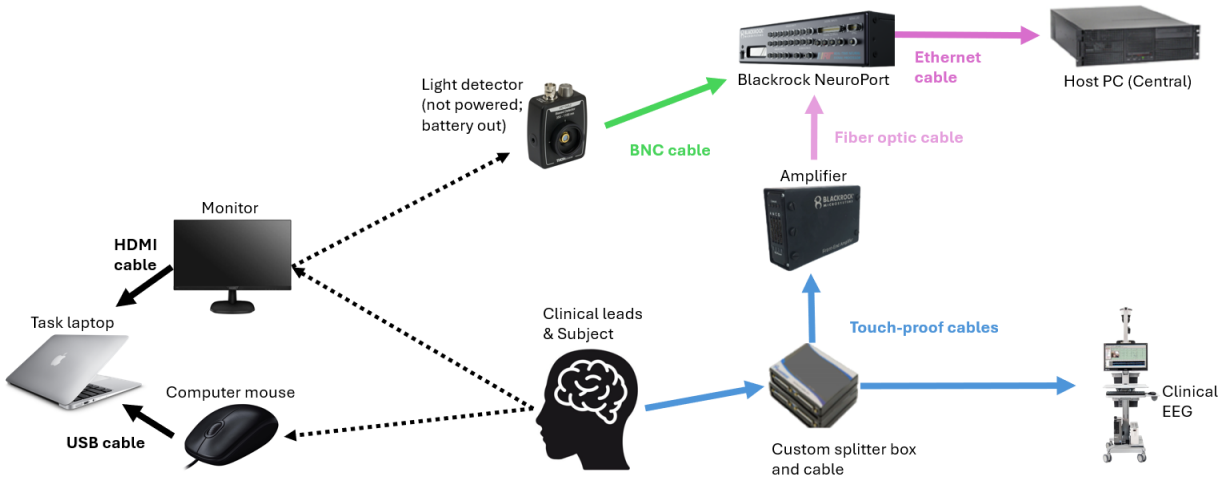


Figure 3.1: **Hardware setup for neural recording and stimulus presentation.** A task laptop presents stimuli on a monitor via HDMI and logs behavioral responses via a USB mouse. A physics lab-grade photodiode (DET36A2) is placed on the monitor to capture precise stimulus onset timing; its analog signal is sent via BNC cable to the Blackrock NeuroPort for synchronization. The subject's implanted SEEG clinical leads connect to a custom splitter box, which passively splits the signals. One path goes to the clinical EEG system for standard hospital monitoring; the other path goes via touch-proof cables to a Blackrock amplifier, then over fiber optic to the NeuroPort, and finally via Ethernet to a central host PC for research recording. This configuration enables identical neural signals to be recorded in parallel by both the clinical system and the research system, time-locked to stimulus onset via the photodiode.

Session Parameters

Recording session parameters are summarized in Table 3.1. Each day's recording session lasted approximately 30 minutes, structured as 3–4 trials of approximately 8 minutes each. Short breaks of a few minutes occurred between trials when needed, but not within trials.

Table 3.1: Recording session parameters for each patient. The "Day" column refers to post-operative day.

Subject	Session	Images	Repeats	Duration (ms)	Delay (ms)
MIT001	day6_session1	50	4	200	300
MIT001	day6_session2	480	4	100	150
MIT001	day6_session3	480	4	100	150
MIT001	day6_session4	480	4	100	150
MIT002	day2_session1	60	2	250-350; uniform	200
MIT002	day2_session2	450	4	125-175; uniform	100
MIT002	day2_session3	480	4	125-175; uniform	100
MIT002	day2_session4	480	4	125-175; uniform	100
MIT002	day5_session1	60	2	125-175; uniform	100
MIT002	day5_session2	480	4	125-175; uniform	100
MIT002	day5_session3	480	4	125-175; uniform	100
MIT002	day5_session4	480	4	125-175; uniform	100
MIT002	day5_session5	NaN	NaN	NaN	NaN
MIT002	day13_session1	480	4	125-175; uniform	100
MIT002	day13_session2	480	4	125-175; uniform	100
MIT002	day13_session3	480	4	125-175; uniform	100
MIT002	day13_session4	480	4	125-175; uniform	100
MIT002	day13_session5	NaN	NaN	NaN	NaN
MIT003	day2_session1	60	2	250-350; uniform	200
MIT003	day2_session2	480	4	125-175; uniform	100
MIT003	day2_session3	480	4	125-175; uniform	100
MIT003	day2_session4	480	4	125-175; uniform	100
MIT003	day5_session1	480	4	125-175; uniform	100
MIT003	day5_session2	480	4	125-175; uniform	100
MIT003	day5_session3	480	4	125-175; uniform	100
MIT003	day7_session1	480	4	125-175; uniform	100
MIT003	day7_session2	480	4	125-175; uniform	100
MIT003	day7_session3	480	4	125-175; uniform	100

3.3 Task Design

Rapid Serial Visual Presentation (RSVP) Paradigm

Patients viewed images presented in a rapid serial visual presentation (RSVP) paradigm while maintaining fixation on a central cross (Figure 3.2). Images were presented sequentially with timing parameters that varied slightly across patients to optimize data collection: (1)

for patient MIT001 fixed image presentation duration of 100 ms with a 150 ms inter-stimulus interval (gray screen), or (2) for patients MIT002 and MIT003, variable image presentation duration drawn uniformly from 125–175 ms with a 100 ms inter-stimulus interval. The temporal jitter was introduced to decorrelate stimulus timing from 60 Hz line noise (since images presented at exactly 4 Hz would have a period that evenly divides the line noise frequency). Both timing protocols yielded qualitatively similar results for visual reliability metrics.

Fixation Cross

A fixation cross was displayed at the center of the screen throughout the experiment, including during both image presentation and inter-stimulus intervals (gray screens). The fixation cross size was 1/64 of the screen dimension.

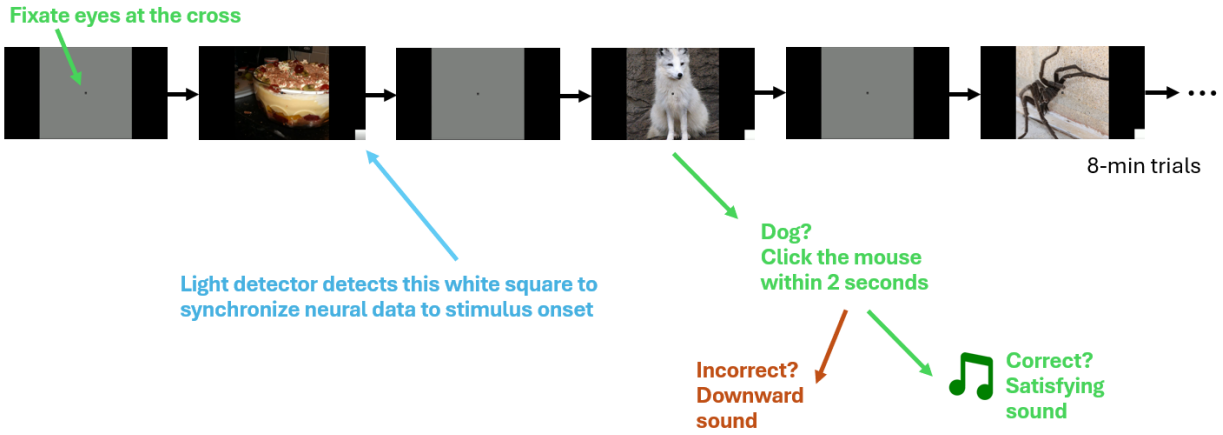


Figure 3.2: **Detailed task paradigm showing the rapid serial visual presentation (RSVP) setup with the auxiliary attention task (dog detection).** Images from multiple datasets are presented sequentially at variable intervals. Patients maintain fixation on the central cross and respond with a mouse click whenever they detect a dog in the images. Correct detections within 2 seconds produce an upward sound; incorrect clicks (no dog in previous 2 seconds) produce a downward sound. At the end of each session, patients receive performance feedback and an in-game medal based on their dog detection accuracy.

Auxiliary Attention Task

To monitor patient attention and increase engagement, we included an auxiliary dog detection task. Patients were instructed to click a mouse whenever they detected a dog in the presented images. Target (dog) images were drawn from ImageNet classes falling under the WordNet 'dog.n.01' synset and from OASIS images with filenames beginning with "dog." If the patient correctly clicked within 2 seconds of a dog image appearing, an upward auditory tone was presented as positive feedback. If the patient clicked without a dog having appeared in the previous 2 seconds, a downward tone was presented as negative feedback.

Approximately 2% of the images contained dogs (i.e. 11 dogs in 480 images). This number was chosen so that a dog would on average appear once per 2-30 seconds, requiring the subject to pay constant attention to the images.

At the end of each session, patients were shown their performance statistics (number of dogs detected out of total dogs presented) and awarded an in-game medal. This gamification was implemented to increase patient engagement and attention throughout the experiment. The auxiliary task served as a verification that patients maintained attention: consistent correct detection of dogs throughout the session indicated sustained engagement. In our experiments, all patients improved at the task over sessions and achieved near-ceiling performance, confirming that the attention intervention was effective.

The auxiliary target images were excluded from the analysis to avoid contamination of the visual reliability measurements with motor or reward confounds etc.

3.4 Visual Stimuli

Image Datasets

Images were drawn from the following three datasets:

- ILSVRC2012 (ImageNet Validation Set): The full ImageNet validation set containing 50,000 images across 1,000 object categories [42]. The validation set was specifically chosen to ensure that no pretrained image models used in subsequent analyses had been trained on these images.
- Open Affective Standardized Image Set (OASIS): A dataset of 900 images spanning various affective and arousal categories [41].
- TrailerFacesHQ: A dataset of 186,553 high-quality face images extracted from movie trailers [43].
- Facial Emotion Recognition (FER) challenge dataset: Only for patient MIT001, we included images from the FER dataset [46] instead of TrailerFacesHQ, consisting of small, low-resolution, grayscale images of faces. This set was subsequently replaced with TrailerFacesHQ in later patients in favor of high-quality color face images.

Image Selection Procedure

Images were selected through a two-stage randomized process without replacement. First, for each frame in the experimental sequence, a source dataset was randomly selected according to a predetermined probability distribution: 70% from ILSVRC2012 (ImageNet validation set), 15% from OASIS, and 15% from TrailerFacesHQ. Second, within the selected dataset, an individual image was randomly chosen from the available pool. For the OASIS dataset, images with blacklisted prefixes (e.g., “nude”) were excluded from the pool.

Target (Dog) Image Classification

Target images (dogs) and non-target images were classified automatically. For the ILSVRC2012 dataset, target images were identified through their ImageNet labels using WordNet's hierarchical semantic structure. Specifically, all images belonging to the 'dog.n.01' synset and its hyponyms were classified as targets, yielding approximately 1,500 dog images from the 50,000-image validation set. For the OASIS dataset, images were classified based on filename prefixes, with all images beginning with "dog" designated as targets. The TrailerFacesHQ dataset contained only non-target images consisting of human faces.

Image Display Parameters

Images were displayed on a monitor with 1920×1080 pixel resolution at 100 Hz refresh rate. Viewing distance was maintained at 70-80 cm, verified at the start of each session, to achieve a visual field of approximately 8-9 degrees of visual angle (corresponding to the approximate extent of high-acuity vision; [47,48]). In one patient (MIT001), we have also experimented with stimuli of size around 16 degrees of visual angle during one of the experimental sessions, and obtained qualitatively similar results. All images were presented in color. Images were cropped to square aspect ratio by cropping the longer dimension, with random crop positioning.

3.5 Signal Processing

Preprocessing

Neural signals were re-referenced using a bipolar montage, computed as the difference between adjacent electrode contacts (for an overview of re-referencing techniques, see [49–53]). No additional filtering was applied to the raw signals. Recording sessions were visually inspected during acquisition using an oscilloscope to verify signal quality.

Frequency Bands

All analyses were performed on raw voltage signals without band-pass filtering. Neural responses were defined according to time windows relative to image onset, as described in the automatic target selection procedure (Section 3.7).

3.6 Split-Half Correlation Analysis

Computing Visual Reliability

We quantified the visual reliability of each neural site using a split-half correlation approach. For each neural site, we computed the split-half correlation coefficient R between response vectors as follows:

Let n denote the number of unique images and m denote the number of repeats per image. For each image, the m repeats were randomly divided into two halves of $m/2$ repeats

each. The neural response for each half was computed by averaging across the $m/2$ repeats within that half. This yielded two n-dimensional vectors (one per half), where each element represents the average response to one image. The split-half correlation R is the Pearson correlation coefficient between these two vectors. The random assignment of repeats to halves was performed independently for each image and was randomized such that the first half of repeats did not necessarily correspond to earlier presentations in the recording session, and all repeats were interleaved randomly throughout the session.

A high split-half correlation indicates that most of the response variance is explained by variability in the visual input, whereas a correlation near zero indicates that neural activity at that site does not reliably depend on the visual stimulus. The underlying assumption is that if the same image is presented twice, the neural response should ideally be the same.

For patient MIT001, in the first recording session we recorded responses to $n = 1,319$ images (excluding the auxiliary target images) with 4 repeats each. For patients MIT002 and MIT003, we recorded responses to $n = 469$ images with 12 repeats each in the first recording session.

Number of Randomizations

For all split-half analyses, we performed 10 random re-orderings of the repeat assignments and averaged the results. Specifically, for each image with m repeats, the order of the m repeats was independently reshuffled before computing each split-half correlation. This averaging helped reduce the variance of the split-half correlation estimate.

Null Distribution

The null distribution was computed by shuffling the image order and all repeats together. Specifically, if there were n images with m repeats each, all $n \times m$ observations were pooled together, shuffled, and then reshaped back into the original shape of the $n \times m$ matrix. This procedure destroys the correspondence between neural responses and image identity, providing a baseline for chance-level correlation under the null hypothesis that the brain response is independent of image content.

3.7 Automatic Target Selection Pipeline

Since there were many parameters to sweep over (100+ neural sites per patient, each with many possible time window choices), we used a systematic parameter sweep to identify optimal targets.

Target Definition

A neural target is defined by the following parameters:

- **Neural site**: An electrode name or pair of electrode names (for bipolar montage).
- **Start time**: The beginning of the response averaging window, specified relative to image onset (e.g., a start time of 0 ms means recording begins at image onset).

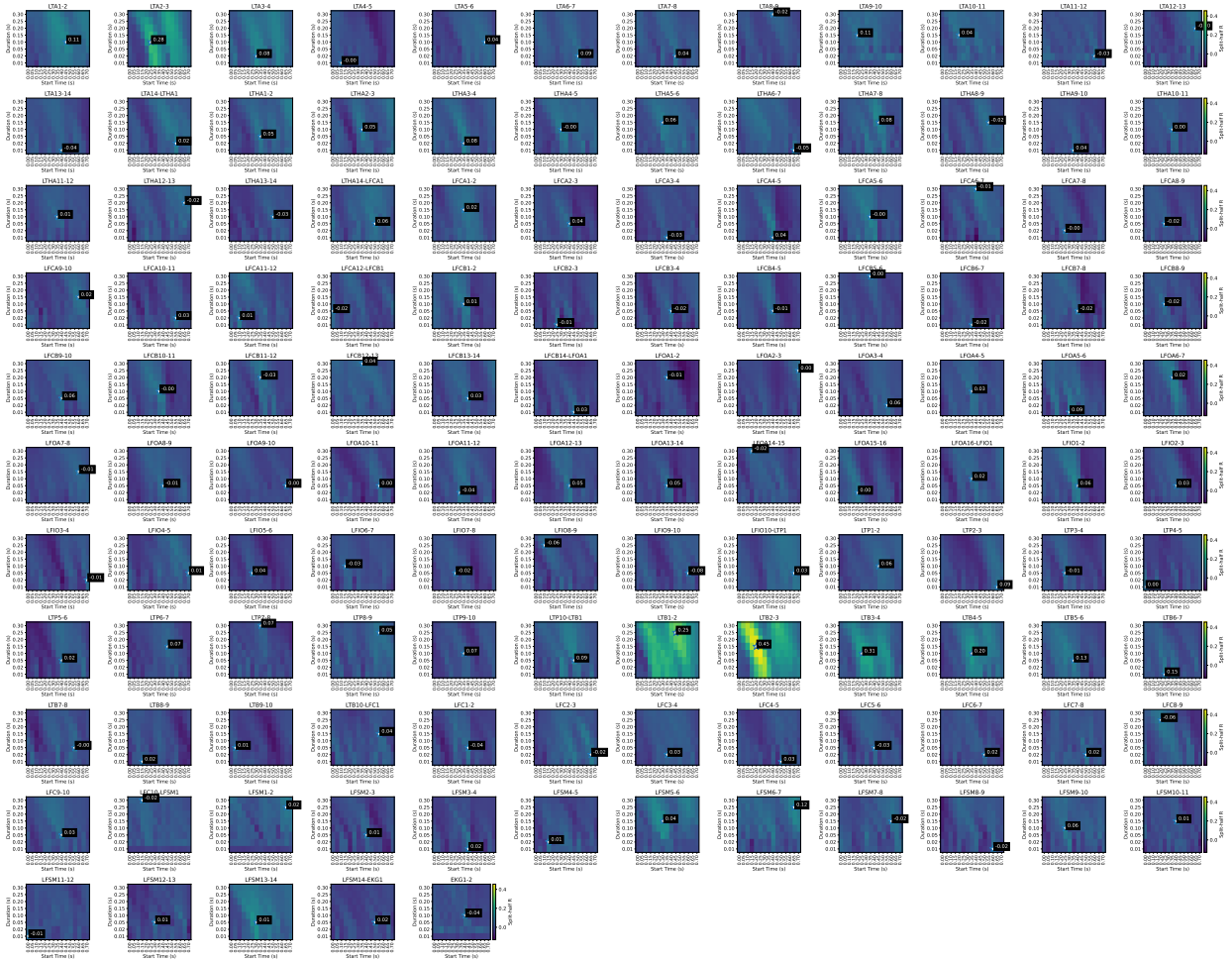


Figure 3.3: **Automatic target extraction pipeline for patient MIT001.** Each panel corresponds to one neural site (bipolar electrode pair). The heatmap shows the split-half correlation coefficient R computed on the training split of images as a function of response window parameters: the x-axis shows the start time of the averaging window relative to image onset, and the y-axis shows the window duration. The star marks the parameter combination yielding the highest training split-half correlation. The text adjacent to each star indicates the test split-half correlation at those optimal parameters, providing an unbiased estimate of visual reliability for that target configuration.

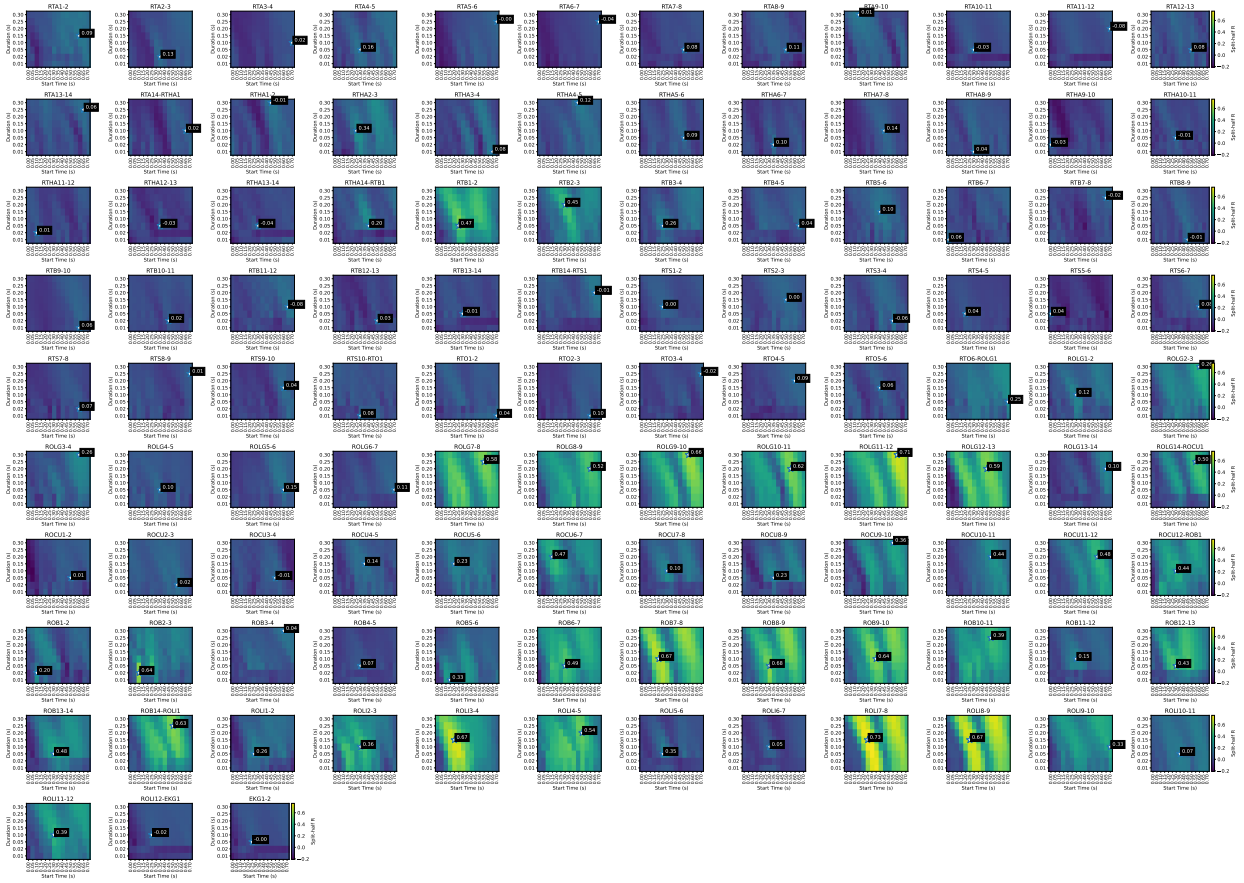


Figure 3.4: **Automatic target extraction pipeline for patient MIT002.** Same format as Figure 3.3.

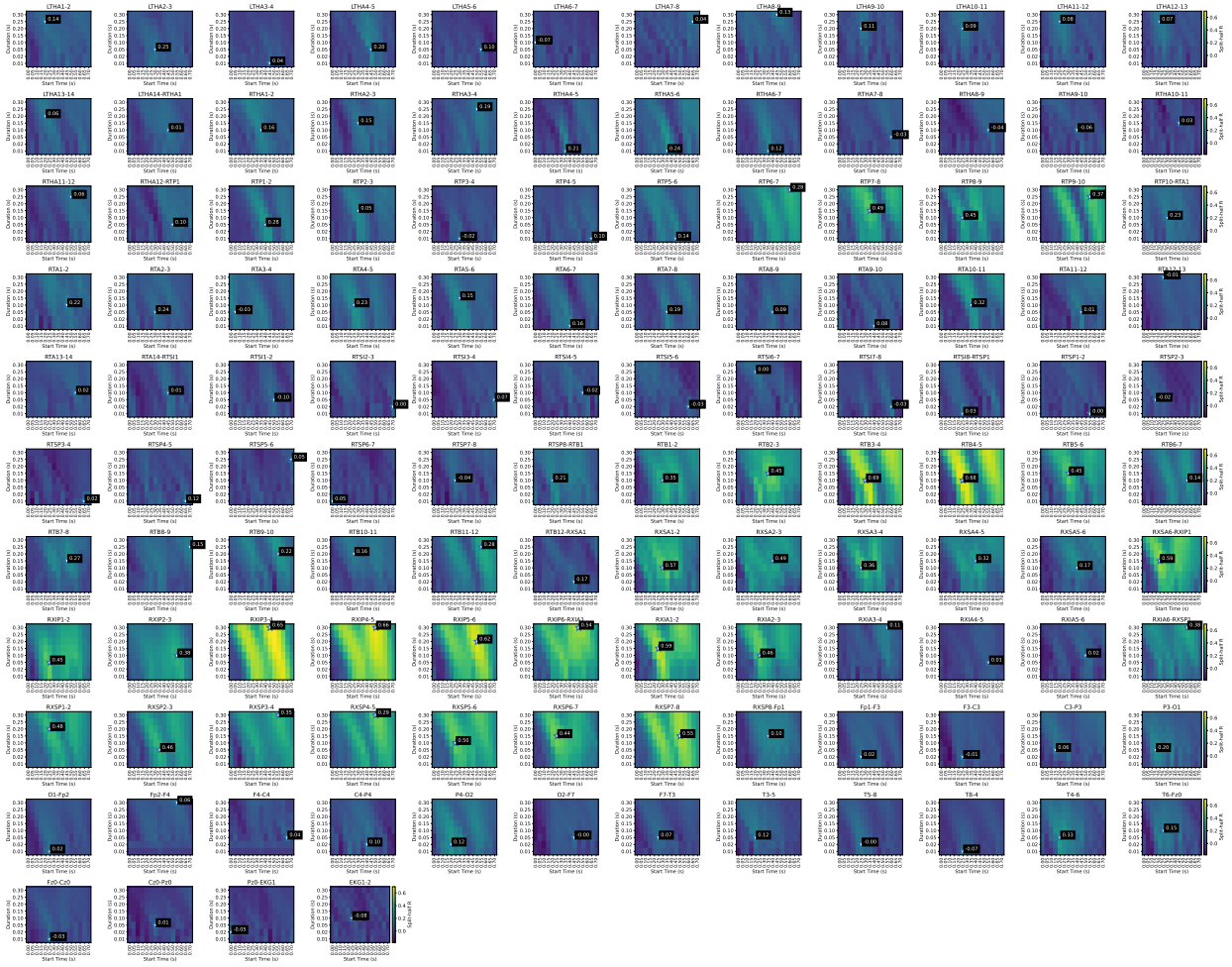


Figure 3.5: **Automatic target extraction pipeline for patient MIT003.** Same format as Figure 3.3.

- **Duration**: The length of the averaging window.

The target response is computed as the mean voltage from the specified electrode within the defined time window.

Parameter Sweep

The sweep parameters were:

- **Start times**: Range: 0.00 to 0.70 seconds relative to image onset, step size: 0.05 seconds
- **Durations**: 0.01, 0.02, 0.05, 0.10, 0.15, 0.20, 0.25, or 0.30 seconds

For each neural site, we computed the split-half correlation for every combination of start time and duration, creating a 2D heatmap of reliability values (Figures 3.3–3.5).

Train-Test Split for Target Selection

To prevent overfitting given the large number of statistical comparisons, we performed the parameter sweep on a training split (first half of all images) and evaluated the selected targets on a withheld test split (second half of all images). The optimal parameters for each neural site were defined as those yielding the highest split-half correlation on the training split. Each target then was tested on the test split to provide an unbiased estimate of target reliability.

Reliable Target Threshold

Targets were classified as "reliable" if they achieved a split-half correlation $R \geq 0.4$ at 4 repeats per image.

3.8 Stability Analysis Over Time

Recording Schedule

Recording days were determined by clinical constraints and patient consent to take part in the experimental task on any given day. Recordings were conducted whenever the patient provided consent and research staff were available. To assess stability across recording days, 55 identical images were presented on all recording days for each patient. These images were used for all cross-day stability analyses. We then quantified it using the Pearson correlation between neural responses to the same images on different days. For each image, the response was averaged across 4 repeats on each day, and the correlation was computed across all 55 stability images.

Stability Metric

Stability was defined as the cross-day correlation divided by the within-day split-half correlation (at 4 repeats). This normalization accounts for the baseline reliability of each target and expresses stability as a proportion of the maximum achievable correlation. For iterative modeling analyses, we defined stable and reliable targets as those that were selected based on the following criteria:

- Minimum reliability threshold: split-half $R \geq 0.4$ on the first recording day
- Minimum stability threshold: cross-day stability ≥ 0.75
- Minimum value of cross-day correlation: $R \geq 0.3$

3.9 Predicting Neural Responses from Image Features

Pretrained Models for Feature Extraction

We extracted image features from the following pretrained deep neural network models:

- Adversarial ResNet-50: ResNet-50, adversarially trained for performance on the ImageNet classification task [54,55]
- Vanilla ResNet-50: Standard ResNet-50, trained on the ImageNet classification task [56].
- CLIP ViT-B/32: The CLIP model with Vision Transformer backbone (ViT-B/32), pretrained in a self-supervised way (OpenAI; [57]).
- DINO ViT-S/16: The DINO model with Vision Transformer backbone (ViT-S/16), pretrained in a self-supervised way (Facebook; [58]).
- ViT-B/16: The Vision Transformer (ViT-B/16), pretrained in a supervised way on the ImageNet classification task (Google; [59]).

Feature Extraction

For each model, features were extracted from the final layer before the softmax classification head. Feature dimensionality varied by model architecture (e.g., 2,048 dimensions for ResNet-50). Features were standardized using z-score normalization (zero mean, unit variance) computed across all images for each feature dimension independently.

Regression Model and Performance Metric

We used ridge regression to predict neural responses from image features. The L2 regularization parameter (λ) was selected via 3-fold cross-validation on the training set using grid search. Training data consisted of all images and all repeats collected on the first recording day. Test data consisted of 4 repeats of randomly selected images (selected using the same probabilistic

procedure as training images) presented for a fixed number of 4 repeats on a subsequent recording day. The same images were never used for both training and testing. Test performance was quantified as R^2 (the square of the Pearson correlation between predicted and observed responses). For population analyses, R^2 values were normalized by the noise ceiling. Error bars represent standard error of the mean across targets.

Noise Ceiling Computation

Suppose on day 1, the neural response is modeled as $f(\text{image}) + \text{noise}_1$, and on day 2 as $g(\text{image}) + \text{noise}_2$, where f and g denote (possibly distinct) signal mappings for each day, and the noise terms are independent.

Let r_{day1} denote the split-half correlation on day 1 (i.e., correlation between two independent noise realizations of $f(\text{image}) + \text{noise}_1$), and r_{12} the cross-day correlation between $f(\text{image}) + \text{noise}_1$ and $g(\text{image}) + \text{noise}_2$ (responses to the same image on day 1 and day 2).

The noise ceiling for a model trained on day 1 and evaluated on day 2, assuming a "perfect" model mapping $f(\text{image})$ to predict $g(\text{image})$, is defined as the model-data correlation:

$$r_{\text{ceiling}} = \frac{\text{Cov}\left(f(\text{image}),\ g(\text{image}) + \text{noise}_2\right)}{\sqrt{\text{Var}(f(\text{image})) \cdot \text{Var}(g(\text{image}) + \text{noise}_2)}}$$

If f and g have variance F and G (over images), and covariance $C = \text{Cov}(f, g)$, then the expected model-data correlation becomes

$$r_{\text{ceiling}} = \frac{C}{\sqrt{F\left(G + \sigma^2_{\text{noise}_2}\right)}}$$

Meanwhile, the split-half reliability on day 1 is:

$$r_{\text{day1}} = \frac{\text{Var}(f(\text{image}))}{\text{Var}(f(\text{image})) + \sigma^2_{\text{noise}_1}} = \frac{F}{F + \sigma^2_{\text{noise}_1}}$$

The cross-day correlation is:

$$r_{12} = \frac{\text{Cov}(f(\text{image}),\ g(\text{image}))}{\sqrt{\left(F + \sigma^2_{\text{noise}_1}\right)\left(G + \sigma^2_{\text{noise}_2}\right)}} = \frac{C}{\sqrt{\left(F + \sigma^2_{\text{noise}_1}\right)\left(G + \sigma^2_{\text{noise}_2}\right)}}$$

Assuming $f = g$ (or equivalently that $C^2 = FG$), and $\sigma^2_{\text{noise}_1} = \sigma^2_{\text{noise}_2}$ (for notational clarity, but the result holds without loss of generality),

$$r_{12} = \frac{F}{F + \sigma^2_{\text{noise}}} = r_{\text{day1}}$$

but if $f \neq g$, in general, the quantities are related by:

$$r^2_{\text{ceiling}} = \frac{C^2}{F(G + \sigma^2_{\text{noise}_2})}$$

and

$$r_{12}^2 = \frac{C^2}{(F + \sigma_{\text{noise}_1}^2)(G + \sigma_{\text{noise}_2}^2)}$$

So,

$$\frac{r_{12}^2}{r_{\text{day1}}} = \frac{C^2}{(F + \sigma_{\text{noise}_1}^2)(G + \sigma_{\text{noise}_2}^2)} \cdot \frac{F + \sigma_{\text{noise}_1}^2}{F} = \frac{C^2}{F(G + \sigma_{\text{noise}_2}^2)} = r_{\text{ceiling}}^2$$

Thus,

$$r_{\text{ceiling}}^2 = \frac{r_{12}^2}{r_{\text{day1}}} \qquad \Longrightarrow \qquad r_{\text{ceiling}} = \sqrt{\frac{r_{12}^2}{r_{\text{day1}}}} \tag{3.1}$$

In summary, the noise ceiling for cross-day prediction is given by $r_{\text{ceiling}} = \sqrt{r_{12}^2 / r_{\text{day1}}}$.

3.10 Scaling Laws Analysis

We systematically varied the training set size by subsampling both the number of unique images (range: 10–400 images) and the number of repeats per image (range: 1–12 repeats). Images were randomly subsampled for each condition. Only data from subjects MIT002 and MIT003 were used for this analysis because they had the most repeats per image. Based on results from the model comparison analysis in this study, the adversarially trained ResNet-50 was selected for scaling law analyses as it achieved the highest R^2 performance. For all scaling conditions, test performance was evaluated on a separate recording day with 4 repeats per image, which provided a consistent benchmark regardless of training configuration. Each scaling condition was repeated with 10 random seeds to estimate variability.

3.11 Driving and Suppressing Neural Activity

Stimulus Selection

For each neural target which was stable and reliable (see Section 3.8), the trained regression model was used to predict responses to all images in the combined dataset (ImageNet validation set, OASIS, and TrailerFacesHQ, sampled in the same proportions as the original recording session: 70% ImageNet, 15% OASIS, 15% TrailerFacesHQ). The 30 images with the highest predicted responses were designated as "drive" stimuli, and the 30 images with the lowest predicted responses were designated as "suppress" stimuli. Drive and suppress images were novel to the patient (never seen during training). Models were trained independently for each target. Additionally, new random images from the same pool were selected for each recording session such that the drive and suppress images constituted at most half of the images presented on that session. Each image was presented 4 times, and the mean response across repeats was computed. Differences between drive, suppress, and random image conditions were assessed using the Mann-Whitney U test.

3.12 Iterative Model Improvement

Iteration Protocol

Models were refined over two iterations. Iteration 1 models were trained on Day 2 data only and tested on Day 5. Iteration 2 models were trained on all available data from Days 2 and 5 (including responses to drive and suppress stimuli from Iteration 1) and tested on Days 7 and 13. For each iteration, models were retrained from scratch using all available data from previous recording sessions. Hyperparameters were re-tuned via cross-validation for each iteration. The stable and reliable targets used for iteration analysis were selected based on meeting all reliability and stability criteria (see Section 3.8) for both the second and last recording days.

Quantification of Prediction Accuracy

Prediction accuracy was quantified as the perpendicular distance from data points to the $y = x$ line in model-prediction vs. measured-response scatter plots. Distance was normalized by dividing by the maximum model prediction or measured response on each day, enabling results to be pooled across targets with different response magnitudes. Differences between iterations were assessed using the Mann-Whitney U test. Each data point represents one condition (drive or suppress) for one target on one day.

3.13 Model Interpretation

To interpret model selectivity, we computed model predictions for all images in the ImageNet validation set (50,000 images) and the TrailerFacesHQ dataset. Images were grouped by their class labels (1,000 ImageNet classes plus one "face" class for TrailerFacesHQ images). ImageNet images retained their original class labels. All TrailerFacesHQ images were assigned to a single "face" category without further subdivision. For each target, we identified the 10 classes with the highest mean predicted responses (most driving) and the 10 classes with the lowest mean predicted responses (most suppressing).

3.14 Software

Analysis Software

All analyses were implemented in Python. MNI coordinates of the electrodes were stored and extracted using Brainstorm (in MATLAB; [60]), which is documented and freely available for download online under the GNU general public license. Deep learning models were implemented in PyTorch and scikit-learn [61,62]. Brain maps were created using the Nilearn package [63].

Stimulus Presentation

Stimuli were presented using a custom Python program built with the Pygame library.



Figure 3.6: **Visual reliability across all neural sites for patient MIT001.** Each panel shows one neural site (bipolar electrode pair). The x-axis shows time relative to image onset (0 ms). The y-axis shows the split-half correlation coefficient R. Red shading indicates time points where $p < 0.001$. This figure extends Figure 2.2 from the main text by showing all recorded electrodes rather than representative examples.

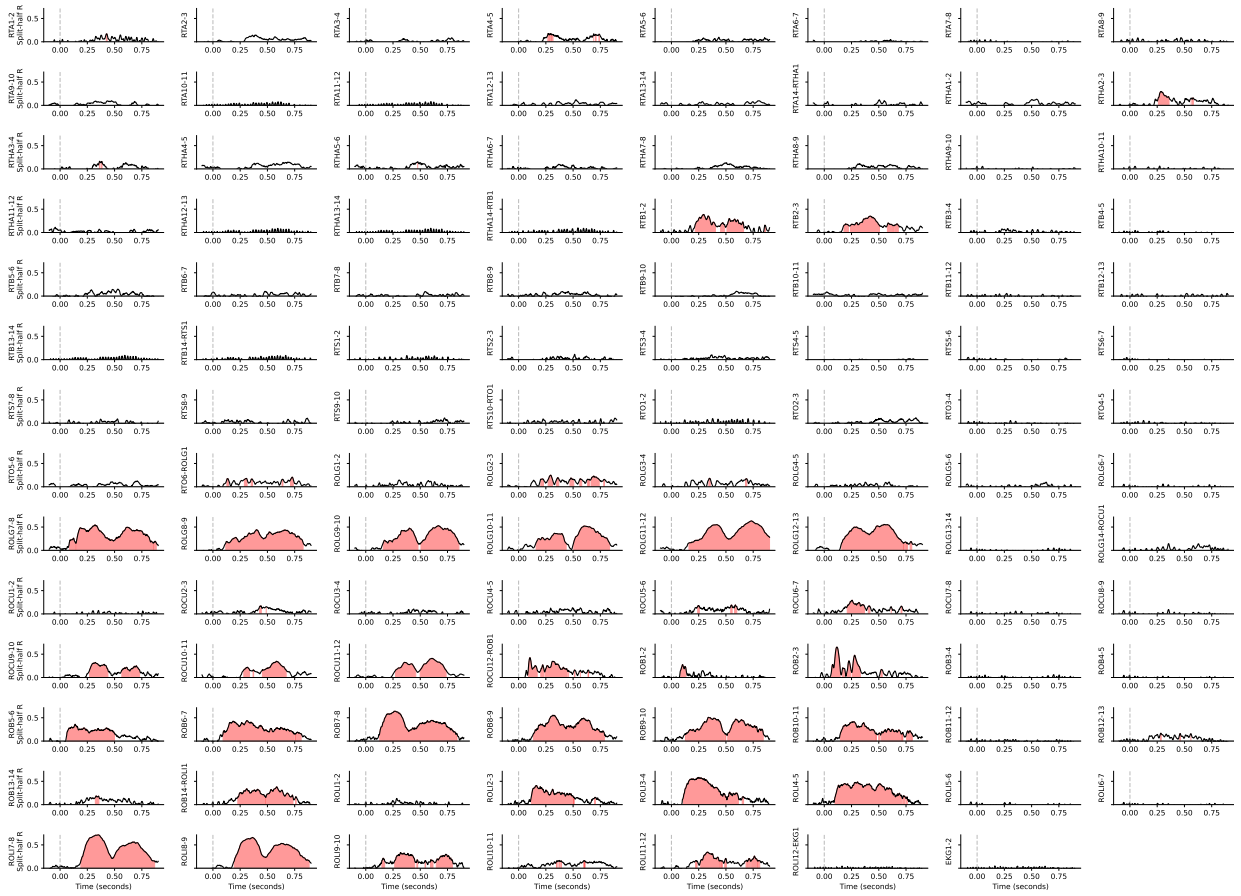


Figure 3.7: **Visual reliability across all neural sites for patient MIT002.** Same format as Figure 3.6.

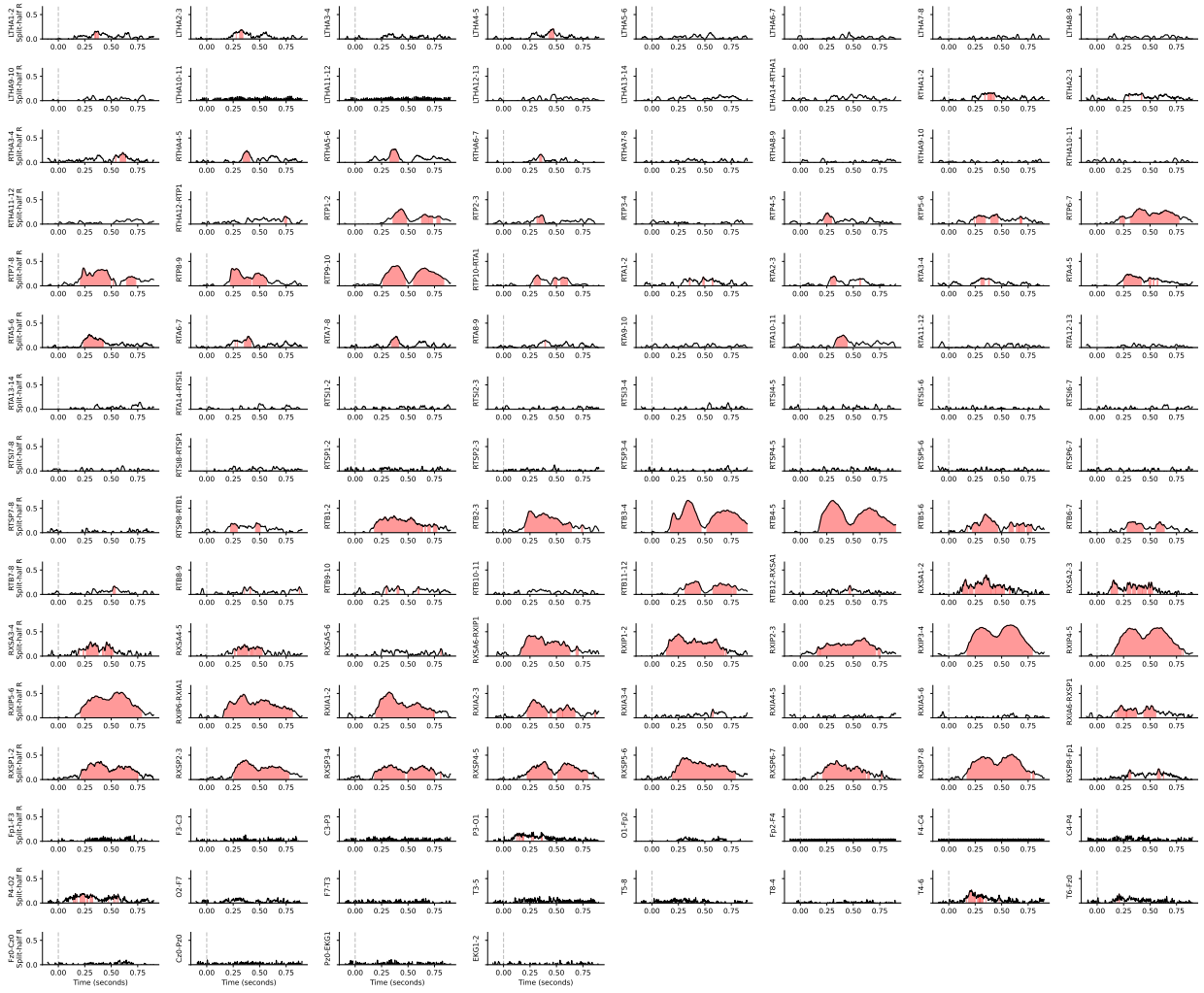


Figure 3.8: **Visual reliability across all neural sites for patient MIT003.** Same format as Figure 3.6.

Chapter 4

Discussion

Summary of Contributions

Using intracranial recordings, this work shows that model-based stimulus optimization can selectively drive and suppress neural activity in the human brain. We show that deep learning models trained on SEEG responses to natural images can predict which stimuli will selectively modulate activity at specific cortical sites, establishing a proof-of-concept for non-invasive neural control through visual sensory stimulation in humans. We established a map of visual reliability across the human brain, introduced the first SEEG dataset with rapid serial visual presentation of natural images, and characterized scaling laws that provide principled guidance for experimental design. These results validate that frameworks successful in non-human primates extend to human intracranial electrophysiology.

Temporal Stability of Neural Targets

A major finding of this work is that neural target stability varies substantially across recording sites and degrades over time for many electrodes. Some targets, such as ROLI7-8 in subject MIT002, maintained high correlations across 13 days (Figure 2.4), while others showed marked degradation. We found a significant correlation between the stability of raw voltage distributions and cross-day image reliability (Figure A.4), suggesting that changes in electrical properties (potentially due to electrode impedance drift, tissue inflammation, scarring, or electrode motion) contribute to the degraded stability. However, we did not yet ascertain the precise mechanisms of this stability degradation. While electrode movement is constrained to less than 1 mm for the clinical care of the patients, even displacements of that amount are large on the scale of neuronal circuits and could affect what is being recorded. Unlike non-human primate protocols, which sometimes instruct researchers to wait multiple weeks post-implantation for electrodes to fully stabilize [64,65], our clinical setup (in the epilepsy monitoring unit) requires recordings to begin within days of surgery and conclude within approximately two weeks. This clinical constraint might also contribute to the stability degradation. This stability issue limits our experimental design and is a challenge for translating this approach to clinical applications. Future work should identify the sources of degraded stability and develop strategies to mitigate them.

Scaling Laws for Experimental Design

Our systematic analysis of scaling laws revealed that test R^2 improves approximately linearly with the logarithm of training set size (see 2.6), which is a relationship consistent with prior work in natural language processing [35,36]. Both the number of unique images and the number of repeats per image contributed to model performance, though for a fixed total budget, adding more unique images yields greater gains than increasing the number of repeats of each image. These scaling relationships provide guidance for experimental design: with limited recording time, researchers should prioritize stimulus diversity over the number of repeats. The scaling curves suggest that model performance has not saturated at our scale, suggesting that increasing the dataset size would yield additional improvements.

Limitations and Future Directions

In this work, we present results from $N = 3$ patients, with only two of them providing multi-day stability data, and our limited electrode coverage is determined entirely by clinical needs and is spatially sparse. Furthermore, due to occipital cortex receiving fewer probes on average for clinical reasons, we lack systematic sampling of visual cortex. Out of more than 200 electrodes, $N = 23$ targets (about 10% of recorded channels) met our reliability and stability criteria for neural control experiments. In the future, we hope to expand our dataset to include a larger number of study participants.

Our framework assumes that ongoing neural activity is driven only by the current visual input, ignoring temporal dynamics and history-dependence. A more complete model would predict neural responses as a function of past neural activity as well as the sensory inputs. However, implementing such dynamics online presents significant technical challenges: stimuli must be synthesized in real-time as neural activity is being recorded, requiring fast model inversion or distillation of large predictive models into compact stimulus generators. Self-supervised pretrained models that could learn to predict neural activity as a function of neural history as well as ongoing sensory inputs are a promising direction for our future work.

We selected stimuli from fixed image datasets, which constrains achievable drive and suppress levels to those of images that are available in the dataset. However, in theory, one could use generative models to synthesize images optimized to elicit responses beyond the range of natural images, potentially achieving stronger neural modulation. In fact, this has been demonstrated in non-human primates [16]. Future work could explore this direction in humans.

We used linear regression on top of frozen pretrained image models such as ResNet-50, CLIP, DINO, and ViT. In the future, we would like to explore whether the use of nonlinear decoders (e.g., small neural networks) instead of the linear decoding head, or fine-tuning the full pretrained model to capture response properties more accurately, would improve predictive performance.

Broader Applications

While this work focused on driving or suppressing individual neural sites, our framework can be extended to eliciting complex spatiotemporal patterns across many neural sites. By

optimizing stimuli to produce desired activity profiles across multiple contacts simultaneously, one could target specific neural computations.

We propose to combine our stimulus optimization framework with models that decode clinically relevant variables from neural activity (for example, seizure prediction models) to identify sensory inputs that optimally modulate those variables. For instance, a model mapping stimuli to neural responses can be composed with a model mapping neural responses to seizure probability. This process yields an end-to-end mapping from stimuli to seizure risk. Inverting this composed model and using it to optimize stimuli could identify visual stimuli that reduce seizure likelihood, opening possibilities for non-invasive therapeutic interventions.

Our approach is not limited to intracranial recordings. In fact, we have shown significant reliability on EEG contacts as well (Figure 3.8). Beyond what has been explored in prior work, we propose that the same framework can be applied to magnetoencephalography (MEG), functional near-infrared spectroscopy (fNIRS), and functional ultrasound imaging (fUSI). Non-invasive modalities enable substantially larger-scale data collection across many more study participants.

Beyond neural control applications, this dataset and modeling framework provide tools for studying visual processing in the human brain at millisecond timescale with direct neural measurements. Unlike fMRI studies, SEEG offers high temporal resolution; unlike EEG, it provides spatially localized signals. Our work validates that computational models of vision generalize from non-human primates to humans, enabling human-specific investigations of visual processing in the brain.

Conclusion

This work establishes that predictive models can guide the selection of visual stimuli to selectively modulate neural activity in the human brain. Our analysis of scaling laws provides quantitative guidance for experimental design, and our framework can be extended to other neural recording modalities and to more complex stimulus optimization objectives (for example, eliciting complex spatiotemporal patterns across multiple neural sites). Taken together, while significant technical and biological challenges remain, this proof-of-concept demonstrates the feasibility of model-guided sensory stimulation as a tool for understanding and ultimately modulating human neural activity.

Appendix A

Appendix: Stability of Targets Across Days

A.1 Intra-day Reliability Remains Stable Across Recording Days

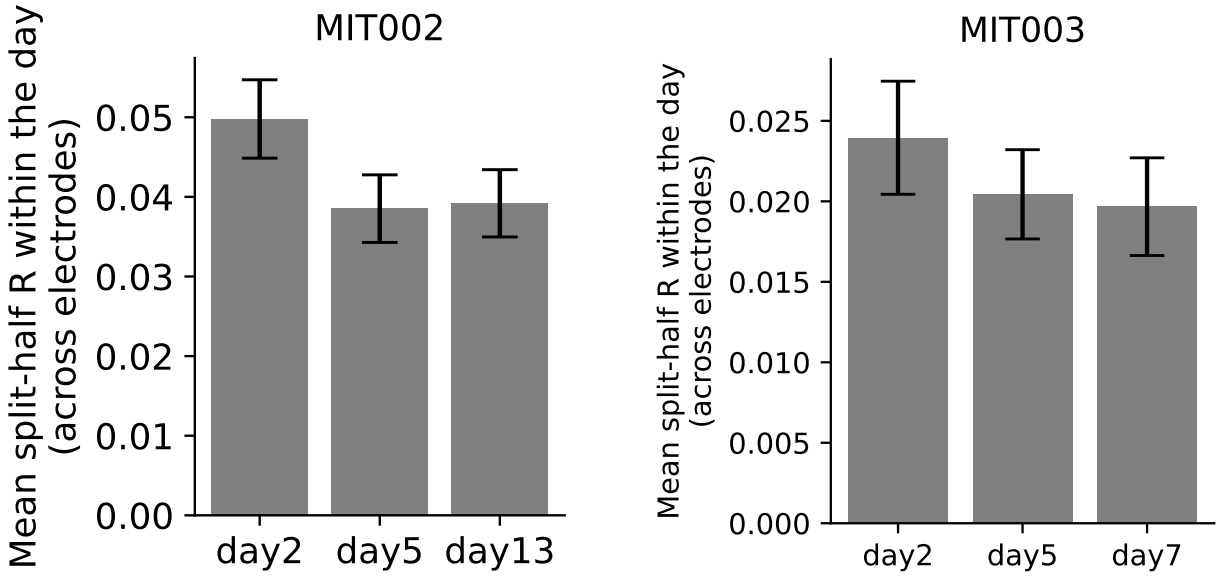
To assess whether visual reliability degrades over time (within the patient's hospital stay), we examined split-half correlation within each recording day. Unlike the cross-day stability analysis (Section 2.2), this measures response consistency within every recording day. For each day, we computed split-half correlation for every neural site across time steps from -0.1 to 0.9 s relative to image onset. Values represent the mean across all time steps; error bars show standard error of the mean across all neural sites (unfiltered, including both visually responsive and non-responsive electrodes).

As shown in Figure A.1, intra-day reliability remains approximately stable for both patients. While reliability appears to slightly decrease, no comparison reached significance (all $p > 0.08$), suggesting the intrinsic capacity for visual encoding does not degrade over the hospital stay timescale.

A.2 Raw Signal Distribution Analysis Across Recording Days

Changes in recording setup electrical properties (amplifier gain, electrode impedance, wiring resistance) could show up as changes in raw voltage distributions, potentially reducing cross-day stability. We examined raw voltage sample distributions from every neural site on each recording day, constructing histograms of all samples (Windsorized at 1% top/bottom). Different days are shown as different colors.

As shown in Figures A.2 and A.3, some electrodes exhibit remarkably stable distributions across days, while others show substantial changes in width or shape, suggesting altered electrical properties. Drive/suppress target electrodes are highlighted in red.



(a) Patient MIT002

(b) Patient MIT003

Figure A.1: **Intra-day visual reliability remains stable across recording days.** Each bar represents the mean split-half correlation coefficient computed within a single recording day, averaged across all time steps (−0.1 to 0.9 s relative to image onset) and across all neural sites (without filtering for reliability). Error bars indicate standard deviation across neural sites. (a) Patient MIT002. (b) Patient MIT003. Neither patient showed statistically significant changes in intra-day reliability across recording days (all pairwise comparisons $p > 0.08$), suggesting that the capacity for visual encoding remains stable throughout the hospital stay.

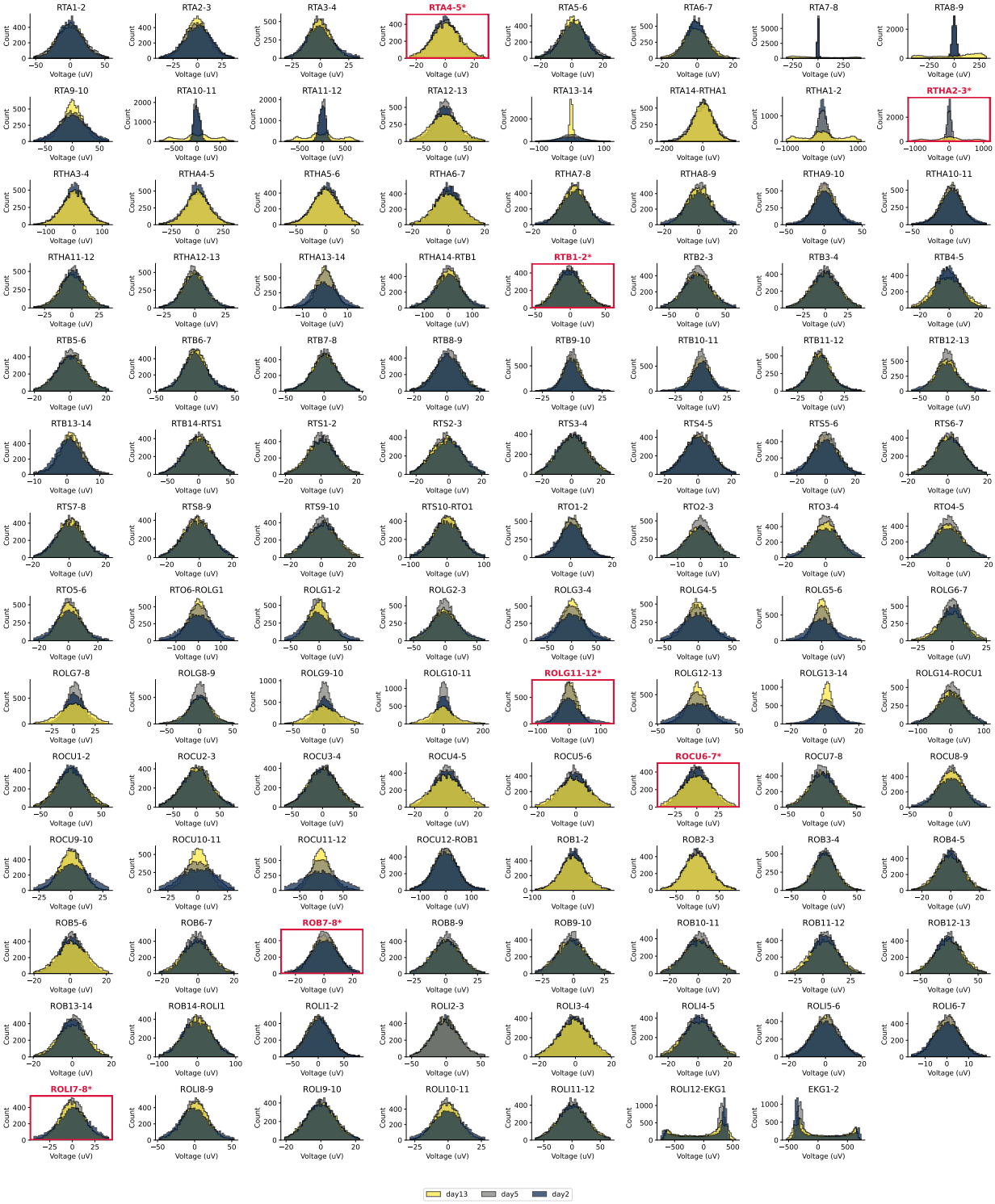


Figure A.2: **Raw voltage sample distributions across recording days for patient MIT002.** Each panel corresponds to one neural site (bipolar electrode pair). Histograms show the distribution of all raw voltage samples collected from each electrode on each recording day, with different colors representing different days. Distributions were Windsorized by removing the top and bottom 1% of samples to reduce the influence of outliers. Some electrodes exhibit remarkably stable distributions across days (nearly overlapping histograms), while others show changes in distribution width or shape. Electrodes that were used as drive/suppress targets in the main analyses are highlighted with red borders.

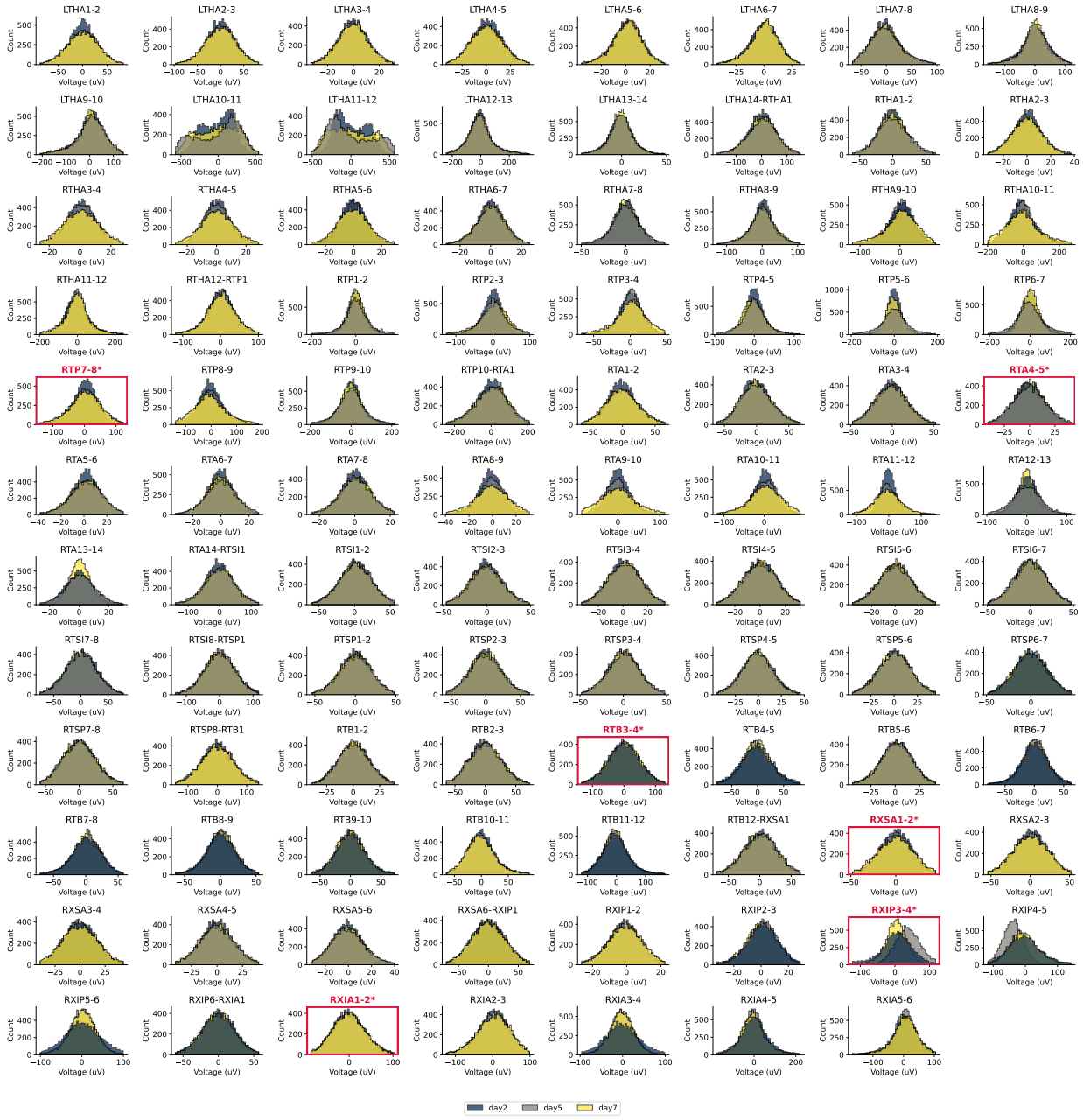
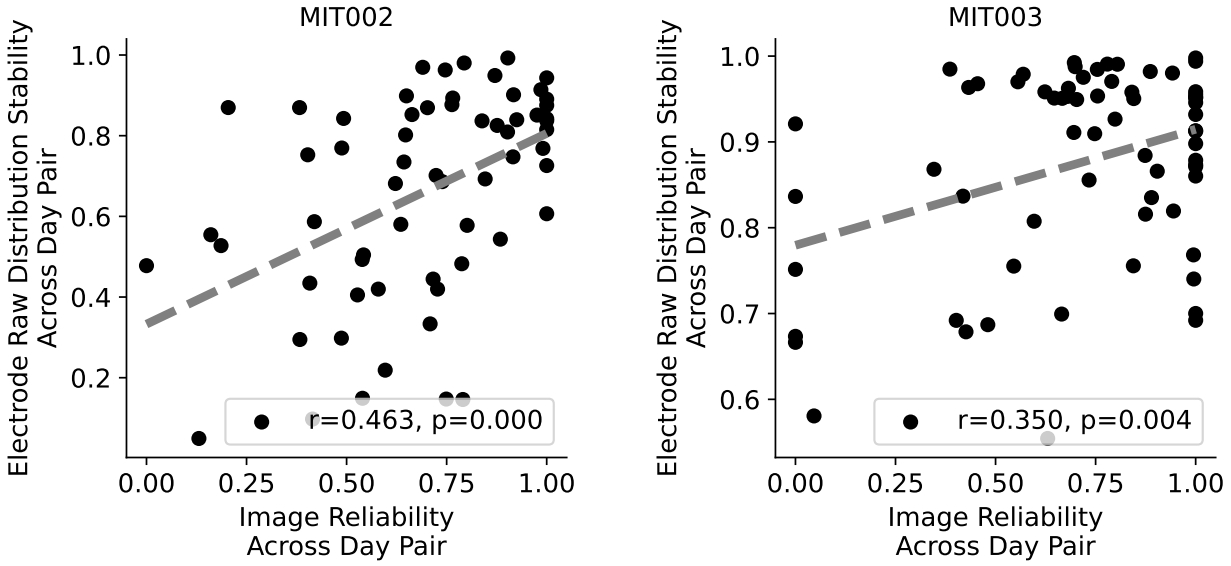


Figure A.3: **Raw voltage distributions for patient MIT003.** Same format as Figure A.2.

A.3 Relationship Between Distribution Stability and Cross-Day Image Stability

To quantify the relationship between stability of the raw voltage distribution and cross-day image stability of neural responses, we considered all pairs of recording days for each electrode. For each day pair, distribution stability was measured as $\min(SD_a/SD_b, SD_b/SD_a)$, which equals 1 for perfectly matched standard deviations and decreases toward 0 as standard deviations diverge. Cross-day image stability for each electrode and day pair was computed as the ratio of the mean split-half correlation across this pair of days to repeated images divided by the mean split-half correlation on the same day, clipped to lie between 0 and 1.

As shown in Figure A.4, a statistically significant positive correlation exists between distribution stability and cross-day reliability in both patients. This suggests that changes in raw electrical properties (for example day-to-day impedance changes, scarring, electrode drift, etc.) could contribute to cross-day reliability degradation.



(a) Patient MIT002

(b) Patient MIT003

Figure A.4: **Relationship between distribution stability and cross-day reliability.** Each point: one electrode for one day pair. X-axis: cross-day stability for images (cross-day correlation for repeated images on different days divided by the split-half correlation on the first recording day). Y-axis: raw acquired samples distribution stability for that target ($\min(SD_a/SD_b, SD_b/SD_a)$). Both patients show significant positive correlation: electrodes with stable signal properties exhibit higher cross-day visual reliability.

Appendix B

Appendix: Advanced, Temporally Precise Trigger Detection

This section describes the methods used to precisely align neural data with each presented image stimulus. This alignment is technically challenging because neural data is recorded on a separate device (Blackrock Neural Signal Processor) that must be electrically isolated from the task laptop and monitor due to the patient safety requirements at the hospital. Despite recording on two independent systems, the downstream analyses require temporal alignment with millisecond resolution.

B.1 Hardware Configuration

The synchronization system operates as follows. The task computer displays images on a monitor, with each image accompanied by a maximally bright white square in the bottom-right corner of the screen. This square appears simultaneously with image onset and disappears when the image is removed (Figure 3.2). A laboratory-grade photodetector (Thorlabs DET36A2) is positioned directly over the square's location on the monitor. The photodetector is a passive device requiring no external power, ensuring minimal signal contamination. Its analog output is transmitted via BNC cable directly to the Neural Signal Processor (NSP), where it is recorded alongside the neural signals (see Figure 3.1).

When each new image appears, the accompanying white square immediately triggers a voltage change in the photodetector, which is then recorded by the NSP. This section describes the signal processing pipeline used to extract precise image onset times from the raw photodetector signal and, critically, to verify that each detected trigger corresponds exactly to the intended image (a requirement for valid downstream machine learning analyses). The trigger detection system must be robust to noise and fully automatic across varying experimental conditions (room lighting, day-to-day variations in photodetector-to-screen distance, etc.). Missing or incorrectly identifying even a single trigger can cause a complete invalidation of the entire downstream analysis: if image-to-neural-data assignments become misaligned, all subsequent assignments will become incorrect. Given that each recording session contains approximately 2,000 image onsets, and each must be identified precisely and in correct temporal order, achieving this automatically despite system noise presents a

technical challenge.

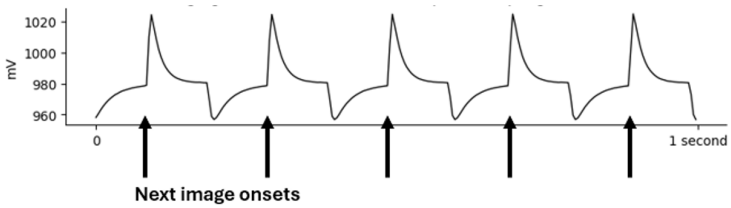


Figure B.1: **Raw photodetector output (y-axis, millivolts) as a function of time (x-axis).** This example displays the photodetector response to five successive image onsets. For each onset, the voltage rises sharply and then decays exponentially.

B.2 Photodetector Response Characteristics

Figure B.1 shows an example time sample of the raw photodetector output with voltage responses to five consecutive image presentations. Each image onset produces a characteristic waveform: a sharp rise followed by an exponential decay. The precise shape of this response function depends on experimental conditions and is not known *a priori*.

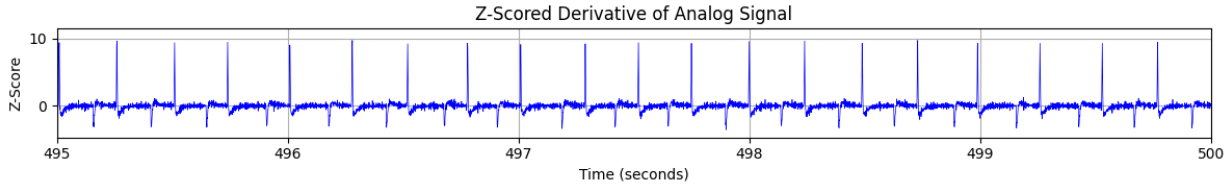


Figure B.2: **Z-scored derivative of the photodetector signal for a representative time segment in an example recording session.** The y-axis shows the z-score of the temporal derivative; the x-axis shows time. Arrows correspond to ongoing image onsets.

B.3 Trigger Detection Pipeline

We developed a multi-stage pipeline to reliably extract trigger times while confirming the accuracy of each detection.

The first stage exploits the fact that the steepest rise of the photodetector signal occurs at the exact moment of image onset. We use this property to obtain an initial set of candidate trigger times, which then allows us to estimate the characteristic response function.

First, we downsample the raw signal by averaging every 10 consecutive samples, reducing the effective sampling rate from 10,000 Hz to 1,000 Hz. We then compute the time derivative by subtracting adjacent values of the signal, followed by z-score normalization. Figure B.2 shows an example result, with spikes appearing at approximately $z = 10$ standard deviations above the mean, corresponding to the sharp rises of the response waveform at each image onset.

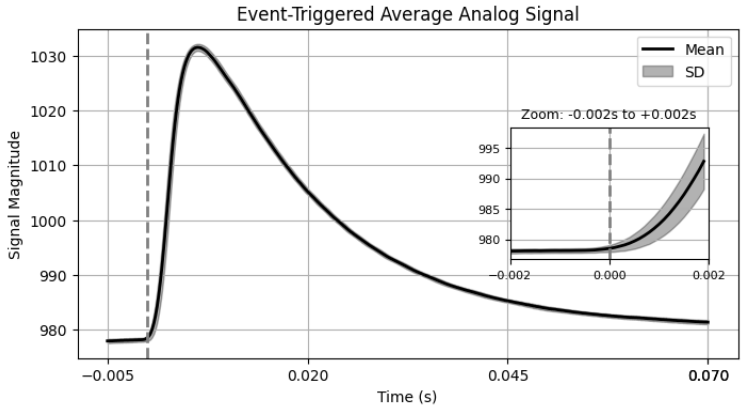


Figure B.3: **Event-triggered average of the photodetector signal, time-locked to high-threshold crossings.** The y-axis shows signal magnitude in millivolts; the x-axis shows time in seconds relative to the threshold crossing. The solid line indicates the mean across all crossings, and the shaded region indicates the standard deviation.

We apply a conservative threshold ($z = 8$) to identify high-confidence threshold crossings. Due to variability across the session's approximately 2,000 total onsets, not all triggers exceed this threshold (typically, around 900 crossings are detected in this first stage), which is sufficient for the waveform estimation purposes.

Using the obtained threshold crossings, we compute an event-triggered average by aligning (time-locking) the raw signal to each crossing and averaging across events. This averaging is performed at the original 10,000 Hz sampling rate to preserve temporal precision for subsequent steps. Figure B.3 shows the resulting template waveform. The inset shows that $t = 0$ corresponds to the sharp onset of the averaged signal. The waveform is manually examined for each recording session. If needed, this alignment can be manually adjusted at this stage. The sharp transition at onset enables temporal precision of approximately ± 1 ms.

The extracted event-triggered average serves as a template for detecting all triggers in the session. We compute the cross-correlation between this template and the full recording at the original 10,000 Hz sampling rate.

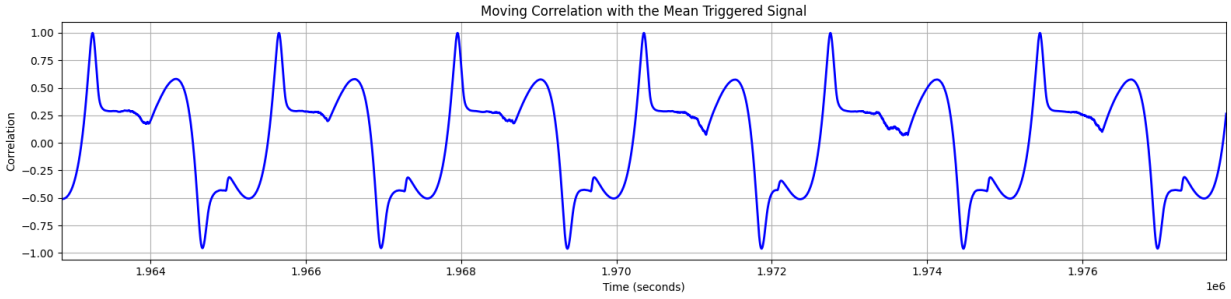


Figure B.4: **Cross-correlation between the template waveform and the full recording session.** The y-axis shows the correlation coefficient; the x-axis shows time. Correlation values approach 1 at times when the recorded signal closely matches the template, corresponding to image onset times.

Figure B.4 shows an example of the resulting moving correlation signal. At times when the recorded signal matches the template, the correlation approaches 1. This method is highly robust to noise because the correlation integrates information across approximately 1,000 data points of the extracted template waveform, effectively averaging out random fluctuations.

We extract the exact number of triggers expected from the task design using a greedy algorithm. The procedure iteratively identifies the maximum correlation value across the session, records that time point as a trigger, blocks out a region around it to prevent re-detection, and continues until the expected number of triggers has been identified. Finally, the detected triggers are sorted temporally to produce the estimated onset time for each image.

B.4 Validation of Trigger Extraction

We perform two complementary validation checks to verify that the detected triggers correspond exactly to the intended images.

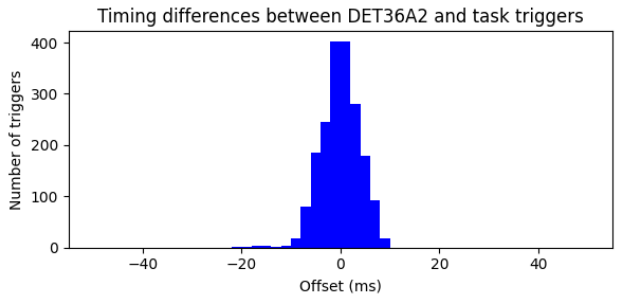


Figure B.5: **Histogram of timing differences between expected task triggers from the task computer and photodetector-extracted triggers.** The y-axis shows the number of triggers; the x-axis shows offset in milliseconds.

The first validation compares the extracted trigger times to the expected times logged by the task computer. Figure B.5 shows a histogram of timing differences for a representative recording session. The distribution is approximately normal, centered near zero, with sharp drop-off around ±10 ms. Since successive images are separated by more than 100 ms in our task design, the tightness of the distribution suggests that our pipeline correctly identifies trigger times. We hypothesize that these ±10 ms offsets between the expected task times and the actual extracted trigger times arise from multiple sources: possible processing delays on the task computer, the discrete 100 Hz monitor refresh rate, delays between image rendering and timestamp logging, and the delays of the system's file I/O.

The second validation ensures not only that trigger times are approximately correct, but that each trigger corresponds exactly to the intended image that is, no triggers are missed (false negatives) and no spurious triggers are detected (false positives). This check exploits the variable inter-stimulus intervals in our task design. We compute the correlation between extracted inter-trigger intervals (from the photodetector pipeline) and expected inter-stimulus intervals (from the task computer). Figure B.6 shows this analysis across 20 time bins spanning the representative example recording session, with correlations computed

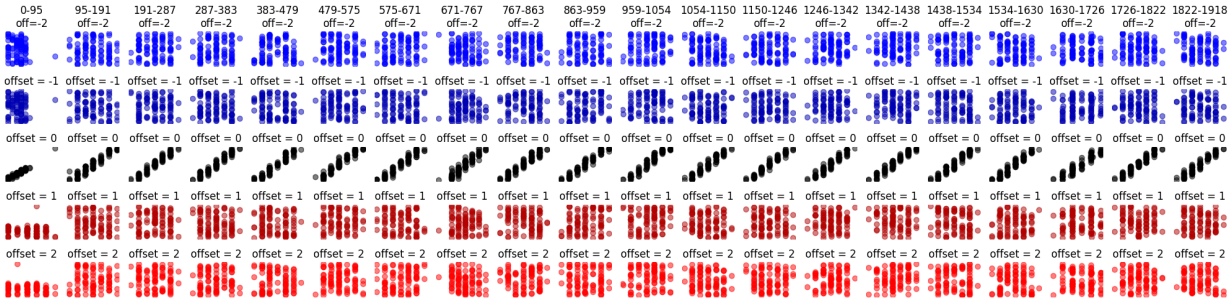


Figure B.6: **Validation of trigger-to-image correspondence.** The figure shows correlations between inter-trigger intervals from the photodetector pipeline and expected inter-stimulus intervals from the task computer. The analysis is performed across 20 time bins (columns), each representing 5% of the recording session. Five rows show correlations at different trigger offsets: -2, -1, 0, $+1$, and $+2$ (blue for negative offsets, black for zero, red for positive).

at trigger offsets of -2, -1, 0, $+1$, and $+2$. At offset $= 0$ (middle row, black), there is a consistent positive correlation across all time bins. This correlation is destroyed at all non-zero offsets (blue for negative, red for positive) throughout the session. This pattern confirms that each detected trigger corresponds exactly to its intended image. If any trigger were missed or spuriously detected, the inter-trigger intervals would become misaligned with expected intervals, and the correlation at offset $= 0$ would become close to zero while correlations at other offsets would show positive values.

Appendix C

Appendix: Extra Figures

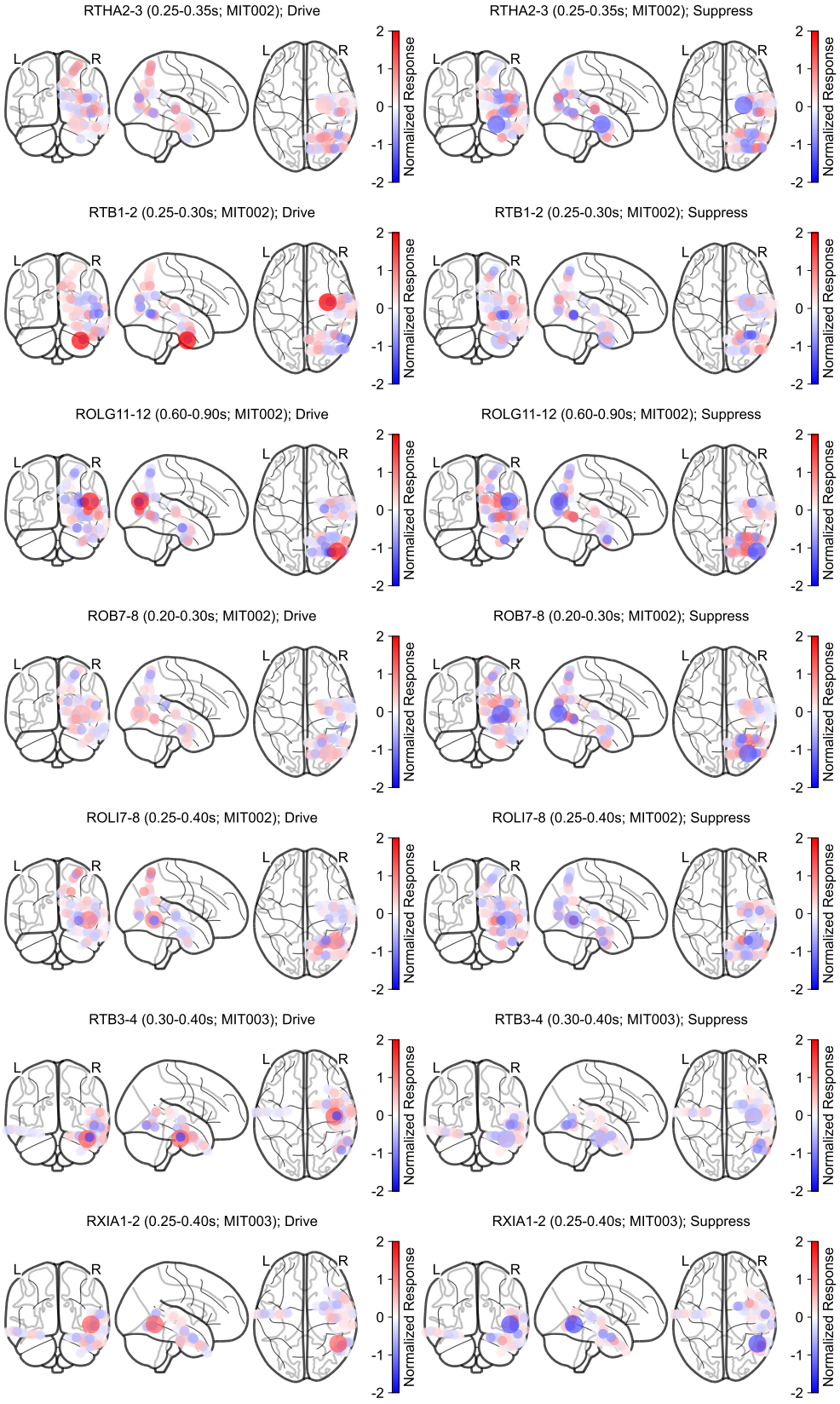


Figure C.1: **Off-target effects of drive and suppress stimuli across the brain.** Each row corresponds to a different neural site being targeted for driving or suppression. Points represent neural sites in an orthographic brain map projection. Color indicates the mean activity of each site in response to the drive/suppress stimuli designed for the target site, normalized by that site's response to random images (z-scored relative to mean and standard deviation). The large dot (2× size) indicates the target site for which stimuli were optimized.

References

[1] S. Sonkusare, M. Breakspear, and C. Guo. "Naturalistic Stimuli in Neuroscience: Critically Acclaimed." *Trends in Cognitive Sciences*, **23**(8), 2019.

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

[3] S.-M. Khaligh-Razavi and N. Kriegeskorte. "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation." *PLOS Computational Biology*, **10**(11), 2014, e1003915. DOI: 10.1371/journal.pcbi.1003915.

[4] M. Schrimpf et al. "Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?" *bioRxiv*, 2018, p. 407007. DOI: 10.1101/407007.

[5] N. Kriegeskorte, M. Mur, and P. Bandettini. "Representational similarity analysis – connecting the branches of systems neuroscience." *Frontiers in Systems Neuroscience*, **2**, 2008, p. 4. DOI: 10.3389/neuro.06.004.2008.

[6] G. Tuckute, A. Sathe, S. Srikant, M. Taliaferro, M. Wang, M. Schrimpf, K. Kay, and E. Fedorenko. "Driving and suppressing the human language network using large language models." *Nature Human Behaviour*, **8**, Mar. 2024, pp. 544–561. DOI: 10.1038/s41562-023-01783-7.

[7] R. Antonello, C. Singh, S. Jain, A. Hsu, J. Gao, B. Yu, and A. Huth. "A generative framework to bridge data-driven models and scientific theories in language neuroscience." *arXiv preprint arXiv:2410.00812*, Oct. 2024. Version 1. DOI: 10.48550/arXiv.2410.00812. URL: https://arxiv.org/abs/2410.00812.

[8] E. A. Hosseini, M. Schrimpf, Y. Zhang, S. Bowman, N. Zaslavsky, and E. Fedorenko. "Artificial Neural Network Language Models Predict Human Brain Responses to Language Even After a Developmentally Realistic Amount of Training." *Neurobiology of Language*, **5**(1), 2024, pp. 43–63. DOI: 10.1162/nol_a_00137.

[9] M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. "Artificial Neural Networks Accurately Predict Language Processing in the Brain." *bioRxiv*, 2020. DOI: 10.1101/2020.06.26.174482.

[10] R. J. Antonello, A. R. Vaidya, and A. G. Huth. "Scaling laws for language encoding models in fMRI." In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 21895–21907.

[11] D. L. K. Yamins and J. J. DiCarlo. "Using goal-driven deep learning models to understand sensory cortex." *Nature Neuroscience*, **19**(3), 2016, pp. 356–365. DOI: 10.1038/nn.4244.

[12] C. R. Ponce, W. Xiao, P. F. Schade, T. S. Hartmann, G. Kreiman, and M. S. Livingstone. "Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences." *Cell*, **177**(4), May 2019, 999–1009.e10. DOI: 10.1016/j.cell.2019.04.005.

[13] E. Y. Walker, F. H. Sinz, E. Cobos, T. Muhammad, E. Froudarakis, P. G. Fahey, A. S. Ecker, J. Reimer, X. Pitkow, and A. S. Tolias. "Inception loops discover what excites neurons most using deep predictive models." *Nature Neuroscience*, **22**(12), 2019, pp. 2060–2065. DOI: 10.1038/s41593-019-0517-x.

[14] P. Bashivan, K. Kar, and J. J. DiCarlo. "Neural population control via deep image synthesis." *Science*, **364**(6439), May 2019, eaav9436. DOI: 10.1126/science.aav9436. URL: https://www.science.org/doi/10.1126/science.aav9436.

[15] C. Guo, M. J. Lee, G. Leclerc, J. Dapello, Y. Rao, A. Madry, and J. J. DiCarlo. "Adversarially trained neural representations may already be as robust as corresponding biological neural representations." *arXiv preprint arXiv:2206.11228*, June 2022. Presented at ICML 2022. DOI: 10.48550/arXiv.2206.11228. arXiv: 2206.11228 [q-bio.NC].

[16] G. Gaziv, S. Goulding, A. Ayvazian-Hancock, Y. Bai, and J. J. DiCarlo. "Noninvasive precision modulation of high-level neural population activity via natural vision perturbations." *arXiv preprint arXiv:2506.05633*, June 2025. Version 3, submitted 14 Jun 2025. DOI: 10.48550/arXiv.2506.05633. arXiv: 2506.05633 [q-bio.NC].

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

[18] A. LeBel, L. Wagner, S. Jain, A. Adhikari-Desai, B. Gupta, A. Morgenthal, J. Tang, L. Xu, and A. G. Huth. "A natural language fMRI dataset for voxelwise encoding models." *Scientific Data*, **10**(1), 2023, p. 555. DOI: 10.1038/s41597-023-02378-9.

[19] S. A. Nastase et al. "The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension." *Scientific Data*, **8**(1), 2021, p. 250. DOI: 10.1038/s41597-021-01033-3.

[20] J. Li, S. Bhattasali, S. Zhang, B. Franzluebbers, W.-M. Luh, R. N. Spreng, J. R. Brennan, Y. Yang, C. Pallier, and J. Hale. "Le Petit Prince multilingual naturalistic fMRI corpus." *Scientific Data*, **9**(1), 2022, p. 530. DOI: 10.1038/s41597-022-01547-9.

[21] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, et al. "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence." *Nature Neuroscience*, **25**(1), 2022, pp. 116–126. DOI: 10.1038/s41593-021-00962-x.

[22] Z. Gong, M. Zhou, Y. Dai, Y. Wen, Y. Liu, and Z. Zhen. "A large-scale fMRI dataset for the visual processing of naturalistic scenes." *Scientific Data*, **10**(1), 2023, p. 559. DOI: 10.1038/s41597-023-02471-x.

[23] M. Visconti di Oleggio Castello, V. Chauhan, J. Guo, and M. I. Gobbini. "An fMRI dataset in response to "The Grand Budapest Hotel", a socially-rich, naturalistic movie." *Scientific Data*, **7**(1), 2020, p. 383. DOI: 10.1038/s41597-020-00735-4.

[24] S. Aliko, J. Huang, F. Gheorghiu, S. Meliss, and J. I. Skipper. "A naturalistic neuroimaging database for understanding the brain using ecological stimuli." *Scientific Data*, **7**(1), 2020, p. 347. DOI: 10.1038/s41597-020-00680-2.

[25] Q. K. Telesford et al. "An open-access dataset of naturalistic viewing using simultaneous EEG-fMRI." *Scientific Data*, **10**(1), 2023, p. 554. DOI: 10.1038/s41597-023-02458-8.

[26] S. Bhattasali, J. Brennan, W.-M. Luh, B. Franzluebbers, and J. Hale. "The Alice Datasets: fMRI & EEG Observations of Natural Language Comprehension." eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by N. Calzolari et al. Marseille, France: European Language Resources Association, May 2020, pp. 120–125. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.15/.

[27] T. Grootswagers, I. Zhou, A. K. Robinson, M. N. Hebart, and T. A. Carlson. "Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams." *Scientific Data*, **9**(1), 2022, p. 3. DOI: 10.1038/s41597-021-01102-7.

[28] W.-L. Zheng and B.-L. Lu. "Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks." *IEEE Transactions on Autonomous Mental Development*, **7**(3), 2015, pp. 162–175. DOI: 10.1109/TAMD.2015.2431497.

[29] N. X. R. Wang, A. Farhadi, R. P. N. Rao, and B. W. Brunton. "AJILE movement prediction: multimodal deep learning for natural human neural recordings and video." In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, 2018. ISBN: 978-1-57735-800-8.

[30] S. M. Peterson, S. H. Singh, B. Dichter, M. Scheid, R. P. N. Rao, and B. W. Brunton. "AJILE12: Long-term naturalistic human intracranial neural recordings and pose." *Scientific Data*, **9**(1), 2022, p. 184. DOI: 10.1038/s41597-022-01280-y.

[31] M. Verwoert, M. C. Ottenhoff, S. Goulis, A. J. Colon, L. Wagner, S. Tousseyn, J. P. van Dijk, P. L. Kubben, and C. Herff. "Dataset of Speech Production in intracranial Electroencephalography." *Scientific Data*, **9**(1), 2022, p. 434. DOI: 10.1038/s41597-022-01542-9.

[32] P. Misra, Y.-C. Shih, H.-Y. Yu, D. Weisholtz, J. R. Madsen, S. Sceillig, and G. Kreiman. "Invariant neural representation of parts of speech in the human brain." *bioRxiv*, 2024. DOI: 10.1101/2024.01.15.575788.

[33] J. Berezutskaya, M. J. Vansteensel, E. J. Aarnoutse, Z. V. Freudenburg, G. Piantoni, M. P. Branco, and N. F. Ramsey. "Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film." *Scientific Data*, **9**(1), 2022, p. 91. DOI: 10.1038/s41597-022-01173-0.

[34] C. Wang et al. "Brain Treebank: Large-scale intracranial recordings from naturalistic language stimuli." In: *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2024. URL: https://openreview.net/forum?id=KZlJF8kguO.

[35] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. *Scaling Laws for Neural Language Models*. 2020. arXiv: 2001.08361 [cs.LG]. URL: https://arxiv.org/abs/2001.08361.

[36] J. W. Rae et al. *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*. 2022. arXiv: 2112.11446 [cs.CL]. URL: https://arxiv.org/abs/2112.11446.

[37] E. Y. Wang et al. "Foundation model of neural activity predicts response to new stimulus types." *Nature*, **640**, Apr. 2025, pp. 470–477. DOI: 10.1038/s41586-025-08829-y. URL: https://doi.org/10.1038/s41586-025-08829-y.

[38] J. Ye et al. "A Generalist Intracortical Motor Decoder." *bioRxiv*, Feb. 2025. Preprint. DOI: 10.1101/2025.02.02.634313. URL: https://doi.org/10.1101/2025.02.02.634313.

[39] D. D. Han et al. *DIVER-1: Deep Integration of Vast Electrophysiological Recordings at Scale*. 2025. arXiv: 2512.19097 [cs.LG]. URL: https://arxiv.org/abs/2512.19097.

[40] Z. Yuan, F. Shen, M. Li, Y. Yu, C. Tan, and Y. Yang. *BrainWave: A Brain Signal Foundation Model for Clinical Applications*. Updated Sep. 2025. 2024. arXiv: 2402.10251 [q-bio.NC]. URL: https://arxiv.org/abs/2402.10251.

[41] B. Kurdi, S. Lozano, and M. R. Banaji. "Introducing the Open Affective Standardized Image Set (OASIS)." *Behavior Research Methods*, **49**(2), Apr. 2017, pp. 457–470. DOI: 10.3758/s13428-016-0715-3.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[43] J. N. M. Pinkney. *Trailer Faces HQ dataset*. https://www.justinpinkney.com/blog/2024/trailer-faces. 2023.

[44] D. J. Kravitz, K. S. Saleem, C. I. Baker, L. G. Ungerleider, and M. Mishkin. "The Ventral Visual Pathway: An Expanded Neural Framework for the Processing of Object Quality." *Trends in Cognitive Sciences*, **17**(1), 2013, pp. 26–49. DOI: 10.1016/j.tics.2012.10.011. URL: https://doi.org/10.1016/j.tics.2012.10.011.

[45] D. J. Felleman and D. C. Van Essen. "Distributed Hierarchical Processing in the Primate Cerebral Cortex." *Cerebral Cortex*, **1**(1), Jan. 1991, pp. 1–47. DOI: 10.1093/cercor/1.1.1-a. URL: https://doi.org/10.1093/cercor/1.1.1-a.

[46] I. J. Goodfellow et al. "Challenges in Representation Learning: A Report on Three Machine Learning Contests." *arXiv preprint arXiv:1307.0414*, 2013. DOI: 10.48550/arXiv.1307.0414.

[47] B. A. Wandell. *Foundations of Vision*. Sunderland, MA: Sinauer Associates, 1995.

[48] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson. "Human Photoreceptor Topography." *Journal of Comparative Neurology*, **292**(4), 1990, pp. 497–523. DOI: 10.1002/cne.902920402. URL: https://doi.org/10.1002/cne.902920402.

[49] T. Mitsuhashi, Y. Iimura, H. Suzuki, T. Ueda, K. Nishioka, K. Nomura, M. Nakajima, H. Sugano, and A. Kondo. "Bipolar and Laplacian montages are suitable for high-gamma modulation language mapping with stereoelectroencephalography." *Frontiers in Neurology*, **15**, 2024, p. 1380644. DOI: 10.3389/fneur.2024.1380644.

[50] J.-P. Lachaux, D. Rudrauf, and P. Kahane. "Intracranial EEG and human brain mapping." *Journal of Physiology-Paris*, **97**(4–6), 2003, pp. 613–628. DOI: 10.1016/j.jphysparis.2004.01.018.

[51] G. Li, S. Jiang, S. E. Paraskevopoulou, M. Wang, Y. Xu, Z. Wu, L. Chen, D. Zhang, and G. Schalk. "Optimal referencing for stereo-electroencephalographic (SEEG) recordings." *NeuroImage*, **183**, 2018, pp. 327–335. DOI: 10.1016/j.neuroimage.2018.08.020.

[52] F. Vidal, B. Burle, L. Spieser, L. Carbonnell, C. Meckler, L. Casini, and T. Hasbroucq. "Linking EEG signals, brain functions and mental operations: Advantages of the Laplacian transformation." *International Journal of Psychophysiology*, **97**(3), 2015. On the benefits of using surface Laplacian (current source density) methodology in electrophysiology, pp. 221–232. ISSN: 0167-8760. DOI: https://doi.org/10.1016/j.ijpsycho.2015.04.022. URL: https://www.sciencedirect.com/science/article/pii/S0167876015001737.

[53] S. Tsuchimoto, S. Shibusawa, S. Iwama, M. Hayashi, K. Okuyama, N. Mizuguchi, K. Kato, and J. Ushiba. "Use of common average reference and large-Laplacian spatial-filters enhances EEG signal-to-noise ratios in intrinsic sensorimotor activity." *Journal of Neuroscience Methods*, **353**, 2021, p. 109089. ISSN: 0165-0270. DOI: https://doi.org/10.1016/j.jneumeth.2021.109089. URL: https://www.sciencedirect.com/science/article/pii/S0165027021000248.

[54] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks." Version 4. *arXiv preprint arXiv:1706.06083*, 2017. ICLR 2018. arXiv: 1706.06083 [stat.ML].

[55] L. Engstrom, A. Ilyas, H. Salman, S. Santurkar, and D. Tsipras. *Robustness (Python Library)*. 2019. URL: https://github.com/MadryLab/robustness.

[56] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[57] A. Radford et al. "Learning Transferable Visual Models From Natural Language Supervision." In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[58] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. "Emerging Properties in Self-Supervised Vision Transformers." In: *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 9630–9640. DOI: 10.1109/ICCV48922.2021.00951. URL: https://doi.org/10.1109/ICCV48922.2021.00951.

[59] A. Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[60] F. Tadel, S. Baillet, J. C. Mosher, D. Pantazis, and R. M. Leahy. "Brainstorm: A User-Friendly Application for MEG/EEG Analysis." *Computational Intelligence and Neuroscience*, **2011**, 2011, p. 879716. DOI: 10.1155/2011/879716.

[61] A. Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[62] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, **12**, 2011, pp. 2825–2830.

[63] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. "Machine learning for neuroimaging with scikit-learn." *Frontiers in Neuroinformatics*, **8**, 2014, p. 14. DOI: 10.3389/fninf.2014.00014. URL: https://doi.org/10.3389/fninf.2014.00014.

[64] J. C. Williams, R. L. Rennaker, and D. R. Kipke. "Long-term neural recording characteristics of wire microelectrode arrays implanted in cerebral cortex." *Brain Research Protocols*, **4**(3), 1999, pp. 303–313. ISSN: 1385-299X. DOI: https://doi.org/10.1016/S1385-299X9900034-3.

[65] C. A. Chestek et al. "Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex." *Journal of Neural Engineering*, **8**(4), 2011, p. 045005. DOI: 10.1088/1741-2560/8/4/045005.