WORKSHOP SUBMISSION: USING MULTIMODAL DNNS TO STUDY VISION-LANGUAGE INTEGRATION IN THE BRAIN

Vighnesh Subramaniam[1,2,*] **Colin Conwell**[3], **Christopher Wang**[1,2],
Gabriel Kreiman[2,4], **Boris Katz**[1,2], **Ignacio Cases**[1,2], **Andrei Barbu**[1,2]
[1]MIT CSAIL [2]CBMM [3]Department of Pscyhology, Harvard University
[4]Boston Children's Hospital, Harvard Medical School
[1]{vsub851,czw,boris,cases,abarbu}@mit.edu
[2]gabriel.kreiman@tch.harvard.edu
[3]conwell@g.harvard.edu

ABSTRACT

We leverage a large stereoelectroencephalography (SEEG) dataset consisting of neural recordings during movie viewing and a battery of unimodal and multimodal deep neural network models (SBERT, BEIT, SIMCLR, CLIP, SLIP) to identify candidate sites of multimodal integration in the human brain. Our data-driven method involves three steps: first, we parse the neural data into discrete, distinct event-structures, i.e., image-text pairs defined either by word onset times or visual scene cuts. We then use the activity generated by these event-structures in our candidate models to predict the activity generated in the brain. Finally, using contrasts between models with or without multimodal learning signals, we isolate those neural arrays driven more by multimodal representations than by unimodal representations. Using this method, we identify a sizable set of candidate neural sites that our model predictions suggest are shaped by multimodality (from 3%-29%, depending on increasingly conservative statistical inclusion criteria). We note a meaningful cluster of these multimodal electrodes in and around the temporoparietal junction, long theorized to be a hub of multimodal integration.

1 INTRODUCTION

The use of deep neural network models to predict and characterize representations in biological cortex is now standard practice in the field of computational cognitive neuroscience. Beginning with seminal work in the primate ventral visual stream (39; 35), this practice has now expanded to include the study of human vision and language cortex alike (36; 18; 19; 27; 8; 12). These studies, however, tend almost uniformly to focus on a single modality of input – vision alone or language alone – in large part because unimodal datasets (9; 2; 4; 30) and unimodal models (e.g. PyTorch-Image-Models; Huggingface) are the most commonly available.

As a product of this unimodal focus, we have learned far less about the correspondence between biological and artificial neural systems tasked with processing visual and linguistic input *simultaneously*. Here, we seek to address this gap by using performant, multimodal deep neural network (DNN) models (VisualBERT, SBERT, BEIT, SimCSE, SIMCLR, CLIP, SLIP) (25; 33; 3; 17; 10; 32; 29) to predict neural activity in a large-scale stereoelectroencephalography (SEEG) dataset consisting of neural responses to the images and scripts of popular movies (38). Our analytic goal is to use systematic comparisons between the neural predictivity of unimodal and multimodal DNNs to identify candidate sites of vision-language integration in the brain. An overview is given in Figure 1(a).

2 METHODS

Neural Data: Invasive intracranial field potential recordings were collected during 21 sessions from 7 subjects (4 male, 3 female; aged $4-19$, $\mu = 11.6$, $\sigma = 4.6$) with pharmacologically intractable epilepsy. During each session, subjects watched a feature length movie from the Aligned Multimodal

[*] Corresponding authors: vsub851@mit.edu, conwell@g.harvard.edu

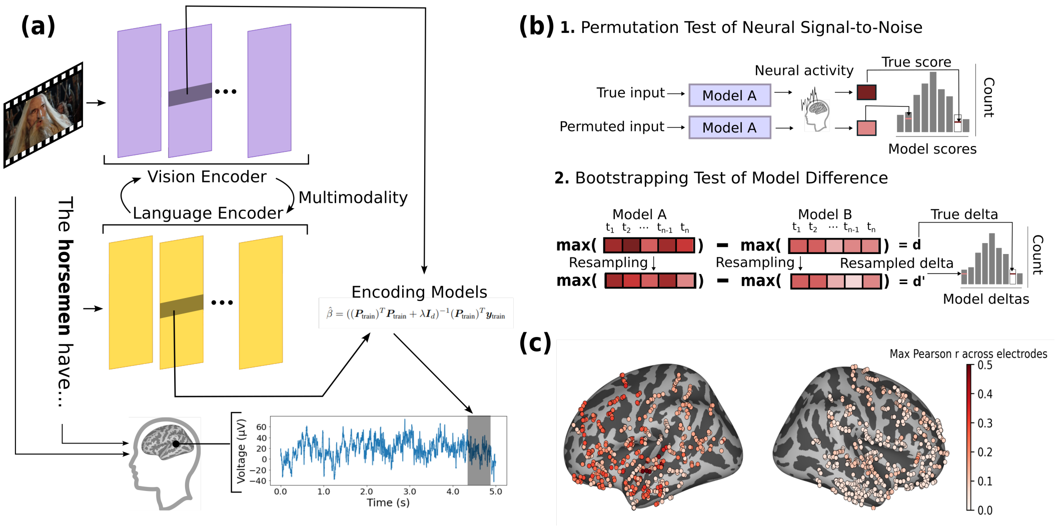

Figure 1: Methods: (a) An overview of our experiment. We use vision, language, and multimodal encoders to predict SEEG activity using a regression. (b) An overview of our analysis to determine the multimodality of a candidate neural site: first, a permutation test to check whether the SEEG signal is meaningfully driven by the vision and language features extracted from the stimulus set; then, a bootstrapping test to compare the performance of two competing models. (Multimodal candidate sites are those sites in which multimodal models are more predictive than unimodal counterparts). (c) Max raw Pearson correlation across our models on all electrodes. Correlation peaks at 0.5 in areas associated with linguistic and visual processing.

Movie Treebank (AMMT) (38) in a quiet room while neural activity was recorded on SEEG probes (28) at a rate of 2kHz. We parse the neural activity into two distinct event-structures where an event structure consists of an image-text pair and create two stimulus alignments where we have aligned visual and language inputs. The first event structure consists of word-onset times, a language-aligned event, and the second consists of visual scene cuts, a vision-aligned event. Word-onset times are collected as part of the AMMT metadata and visual scene cuts are extracted from each movie using PySceneDetect (7). Following (18), we extract a 4000ms window of activity (about 8000 samples), 2000ms prior to the event occurrence and 2000ms after the event occurrence, per electrode. We split the 4000ms window into sub-windows of 200ms with a sliding window of 25ms and the activity is averaged per sub-window to get a series of averaged activity values over time per electrode. A more thorough explanation of our neural data processing can be found in Section B.

Models: We use 9 pretrained deep neural network models, 5 multimodal and 4 unimodal, to explore the effect of multimodality on predictions of neural activity. The models that serve as our main experimental contrast are the SLIP models (29). The SLIP models are a series of 3 models that use the same architecture (ViT-[S,B,L]) and the same training dataset (YFCC15M), but are trained with one of three objective functions: pure unimodal SimCLR-style (10) visual contrastive learning (henceforth SLIP-SimCLR); pure multimodal CLIP-style (32) vision-language alignment (henceforth SLIP-CLIP); and combined visual contrastive learning with multimodal CLIP-style vision-language alignment (henceforth SLIP-Combo). The full set constitutes a set of 5 models (SLIP-SimCLR; the SLIP-CLIP visual encoder; the SLIP-CLIP language encoder; the SLIP-Combo visual encoder; the SLIP-combo language encoder). For more general (uncontrolled) multimodal-unimodal contrasts, we include the mulitmodal model VisualBERT (25) and the unimodal models SBERT (34), BEIT (3), and SimCSE (17). For each of the 9 networks, we assess both a pretrained and randomly-initialized version to assess whether the multimodality we assume in the brain coincides with some form of multimodal learning pressure. (More details on the logic of these choices are given in Appendix A.)

Neural Regression: To identify candidate multimodal areas, we first extract feature vectors from every layer of our candidate networks. We then use these features as predictors in a 5-fold ridge regression predicting the averaged neural activity of a target neural site in response to each event structure. We measure the strength of our regression using the Pearson correlation coefficient between predicted average activity and actual average activity for a *specific* time window in each neural site for a held-out test set of event structures. Two aspects of this process are worth emphasizing: First, our final performance metric (the Pearson correlation between actual and predicted neural activity for a held-out test set of event-structures) is not a correlation over time-series (for which the Pearson correlation is inappropriate), but a correlation over a set of (nominally IID) event-structures that we have extracted by design to minimize the autoregressive confounds of time-series data. Second, our cross-validation procedure and train-test splitting is specifically designed to assess the generalization of our neural regression fits, and as such contains no cross-contamination of selection procedures (e.g. the maximally predictive layer from a candidate DNN, feature normalization, or the ridge regression lambda parameter) and final model scoring. (More details on our regression can be seen in Appendix C.)

Signal versus Noise: Typical neural recording datasets leverage multiple repetitions of the same stimulus to establish various forms of signal-to-noise ratio that indicate whether activity is meaningfully driven by extrinsic differences in stimulus presentation. Given the lack of repetition in our SEEG dataset, we determine whether neural activity is meaningfully driven by our parsed event structures using a permutation test (scrambling the event-structures in each permutation). If across 1000 permutations we find the difference between the permuted and unpermuted score to be greater than 0 in at least 950 cases, we consider the neural activity to be driven by the difference in our event structures. To correct for multiple comparisons, we adjust the resultant p-value for each test in each electrode using standard FDR (Benjamini-Hochberg) corrections (37).

Multimodality Tests: Taking inspiration from fMRI searchlight analyses (23; 14), we perform an additional series of statistical tests on each electrode to determine whether or not they are better predicted by multimodal or unimodal representations. Each test at its core consists of comparing two models with a bootstrapping analysis of their max decoding accuracy across the sliding time windows, with the criteria for being labeled as a 'multimodal' electrode increasingly stringent across tests. The bootstrapping analysis allows us to determine whether the difference in scores between the models is statistically significant with resampling across the sliding time windows as shown in Figure 1(b). We repeat the bootstrapping procedure for all electrodes and use FDR (Benjamni-Hochberg) multiple comparisons corrections to adjust the p-value associated with each electrode on each test. The multimodality logic we apply (in order of stringency) is as follows: (1) Is any multimodal model significantly more predictive than all other unimodal models in *either* of our dataset alignments (word onset, scene cuts)? (2) Is the SLIP-Combo vision transformer significantly more predictive than the SLIP-SimCLR vision transformer in *either* of our dataset alignments? (3) Is any multimodal model significantly more predictive than all other unimodal models in BOTH of our dataset alignments? (4) Is the SLIP-Combo vision transformer more predictive than SLIP-SimCLR vision transformer in BOTH of our alignments? (A more detailed description is given in Appendix D).

3 Results

While there is no single meaningful measure of overall modeling performance, since we expect significant variance in performance as a function of *multiple* controlled and uncontrolled sources, there are a few key metrics we can consider to provide an overall gestalt of our model-to-brain encoding pipeline and the specific measured effects. Unless otherwise noted, we use the following convention in the reporting of these metrics: arithmetic mean [lower 95% confidence interval; upper 95% confidence interval].

As an initial heuristic, we consider the bootstrapped average, as well as the bootstrapped upper and lower bounds on performance across all N = 18 models (9 architectures, with both trained and randomly-initialized weights), N = 2 dataset alignments (word onsets, scene cuts) and all N = 1090 electrodes, after we've selected the max accuracy across time. This constitutes a total of 18 * 2 * 1090 = 39,420 data points. The bootstrapped global average (i.e. the bootstrapped mean) across these data points is r_{Pearson} = 0.0776 [0.0770, 0.0781]. The bootstrapped upper bound (i.e. the bootstrapped max) across these data points is r_{Pearson} = 0.502 [0.515, 0.517]. And the bootstrapped lower bound (i.e. the bootstrapped minimum) is r_{Pearson} = -0.102 [-0.093, -0.0638]. (Negatives here mean model predictions were anticorrelated with ground truth.) This is of course a coarse metric, meant only to give some sense of the encoding performance overall, and to demonstrate its notable range across electrodes. (An illustration of this range is available in Figure 1(c)).

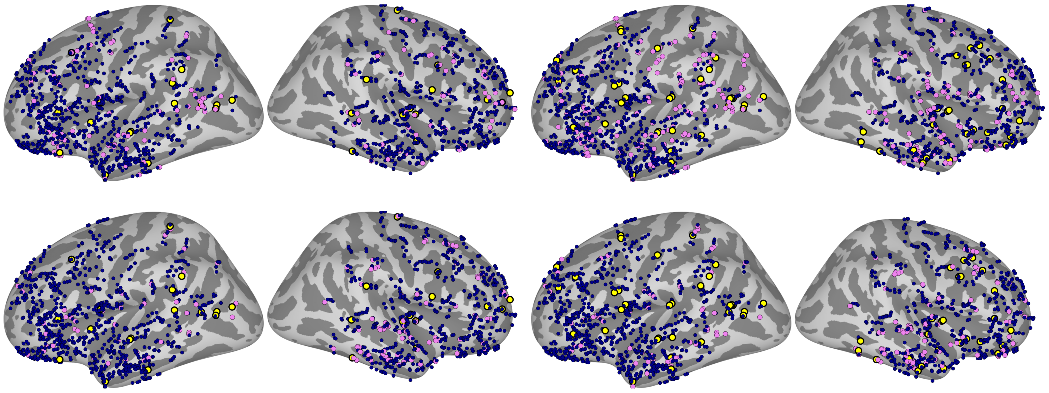

Figure 2: Overview of identified multimodal candidate sites in the brain. The top row compares any multimodal model with any unimodal model and the bottom row compares SLIP Vision encoder with SimCLR. The columns show activity alignment. We highlight all electrodes that are considered multimodal (passing test 1 and 2) in purple. We highlight all electrodes that are considered multimodal in both alignments (passing test 3 and 4) in yellow.

3.1 Multimodality Tests

This brings us to the first of our statistical tests: the permutation test that serves as an indicator of whether our models predict meaningful signal. We find that the vast majority of electrodes in this dataset seem to yield meaningful variance across extracted model features: on average (across the N = 18 models), 989 [967, 1008] / 1090 electrodes pass the permutation test for language-aligned structures (with an average difference of r_{Pearson} = 0.091 [0.086, 0.095] over permuted structures); 894 [872, 913] pass in vision-aligned structures (with a mean gain = 0.119 [0.116, 0.121]). We exclude any model-electrode combination that fails its permutation test from subsequent multimodality tests.

Our first test of multimodality (a filter that selects only those electrodes with a multimodal model significantly more predictive than any unimodal model) yields $323/1090$ electrodes (29.6%) using language-aligned event-structures, and $221/1090$ (20.3%) using vision-aligned event-structures. The average difference in performance between the multimodal model and the next best unimodal model across the 297 language-aligned electrodes was $r_{\text{Pearson}} = 0.019$ $[0.018, 0.022]$; the average difference in vision-aligned electrodes was 0.016 $[0.014, 0.018]$.

Our second test of multimodality (a filter that selects electrodes in which the multimodal SLIP-combo vision transformer significantly outperforms the unimodal SLIP-SimCLR transformer) yields $195/1090$ electrodes (17.9%) using language-aligned event structure, and $181/1090$ (20.2%) using vision-aligned structures. The average performance difference between the SLIP-SimCLR and SLIP-combo vision transformers in language-aligned electrodes was 0.210 $[0.0186, 0.0233]$; the average difference in vision-aligned electrodes was 0.0180 $[0.0153, 0.0208]$.

Our third test of multimodality (a filter that selects only those electrodes with a multimodal model significantly more predictive than a unimodal model in BOTH dataset alignments) yields $73/1090$ electrodes (6.70%). The average difference in performance between the multimodal model and the next best unimodal model across the 73 electrodes was 0.0177 $[0.0153, 0.0203]$.

Our final test of multimodality (a filter that selects only those electrodes in which the multimodal SLIP-combo vision transformer significantly outperforms the unimodal SLIP-SimCLR vision transformer in BOTH dataset alignments) yields $32/1090$ electrodes (2.94%). The average difference in performance between the SLIP-combo vision transformer and unimodal SLIP-SimCLR vision transformer was 0.0215 $[0.0171, 0.0263]$.

Visually inspecting the location of the candidate multimodal sites in Figure 2, we find that the largest contiguous cluster of these electrodes is found in and around the temporoparietal junction. $12/32$ electrodes that pass our most stringent multimodality test (the superiority of the multimodal SLIP-combo vision encoder over its unimodal SLIP-SimCLR counterpart) fall in this junction. In this junction, the superior temporal cortex and middle temporal cortex are commonly associated with language and auditory processing (16; 15) and the inferior parietal lobe is commonly associated with social cognition and human interaction (5). The inherently multimodal abstractions at this junction may be a core reason we find them to be better predicted by multimodal representations.

4 Conclusion

Human intelligence is predicated in large part on the ability to transform sense-perceptual representations into meaningful linguistic tokens, and vice versa. Despite progress, the precise neural correlates of this multimodal, vision-language transformation remain relatively uncharted. The recent ascendance of multimodal deep neural network models could afford us a more direct means of mapping *where and when* this transformation is likely to occur in the brain. Furthermore, these same models might also allow us to more directly investigate *how* this integration occurs. Having access to the learned internal representations of multimodal models means we should be able to perturb these representations in systematic ways and assess the impact of those perturbations on brain predictivity. The candidate sites we've identified in this analysis are still just candidates, but they are precisely the kind of sites we intend to more actively probe in future analyses, addressing the various shortcomings of the current data-driven, correlational approach with more theory-driven, pseudo-causal approaches.

Acknowledgements

This work was supported by the Center for Brains, Minds, and Machines, NSF STC award CCF-1231216, the NSF award 2124052, the MIT CSAIL Systems that Learn Initiative, the MIT CSAIL Machine Learning Applications Initiative, the MIT-IBM Watson AI Lab, the CBMM-Siemens Graduate Fellowship, the DARPA Artificial Social Intelligence for Successful Teams (ASIST) program, the DARPA Knowledge Management at Scale and Speed (KMASS) program, the United States Air Force Research Laboratory and United States Air Force Artificial Intelligence Accelerator under Cooperative Agreement Number FA8750-19-2-1000, the Air Force Office of Scientific Research (AFOSR) under award number FA9550-21-1-0014, and the Office of Naval Research under award number N00014-20-1-2589 and award number N00014-20-1-2643. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 274–281, 2001.

[2] Emily Jean Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Logan T Dowdle, Bradley Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, et al. A massive 7t fmri dataset to bridge cognitive and computational neuroscience. *bioRxiv*, 2021.

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[4] Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. The alice datasets: fmri & eeg observations of natural language comprehension. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 120–125, 2020.

[5] Danilo Bzdok, Gesa Hartwigsen, Andrew Reid, Angela R Laird, Peter T Fox, and Simon B Eickhoff. Left inferior parietal lobe engagement in social cognition and language. *Neuroscience & Biobehavioral Reviews*, 68:319–334, 2016.

[6] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

[7] Brandon Castellano. PySceneDetect v0.6.1, 2022. URL `https://github.com/Breakthrough/PySceneDetect`.

[8] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):1–10, 2022.

[9] Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6 (1):1–18, 2019.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

[11] Colin Conwell, David Mayo, Andrei Barbu, Michael Buice, George Alvarez, and Boris Katz. Neural regression, representational similarity, model zoology & neural taskonomy at scale in rodent visual cortex. *Advances in Neural Information Processing Systems*, 34:5590–5607, 2021.

[12] Colin Conwell, Jacob S Prince, George Alvarez, and Talia Konkle. Large-scale benchmarking of diverse artificial vision models in prediction of 7t human neuroimaging data. *bioRxiv*, 2022.

[13] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[14] Joset A Etzel, Jeffrey M Zacks, and Todd S Braver. Searchlight analysis: promise, pitfalls, and potential. *Neuroimage*, 78:261–269, 2013.

[15] Angela D Friederici. The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences*, 16(5):262–268, 2012.

[16] Angela D Friederici, Michiru Makuuchi, and Jörg Bahlmann. The role of the posterior superior temporal cortex in sentence comprehension. *Neuroreport*, 20(6):563–568, 2009.

[17] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

[18] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. *BioRxiv*, pp. 2020–12, 2021.

[19] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

[20] Trevor Hastie and Robert Tibshirani. Efficient quadratic regularization for expression arrays. *Biostatistics*, 5(3):329–340, 2004.

[21] William B Johnson. Extensions of lipschitz mappings into a hilbert space. *Contemp. Math.*, 26: 189–206, 1984.

[22] Arno Klein and Jason Tourville. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience*, 6:171, 2012.

[23] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868, 2006.

[24] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4953–4963, 2022.

[25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[26] Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 287–296, 2006.

[27] Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031, 2021.

[28] Hesheng Liu, Yigal Agam, Joseph R Madsen, and Gabriel Kreiman. Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, 62(2):281–290, 2009.

[29] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

[30] Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J Honey, Yaara Yeshurun, Mor Regev, et al. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific data*, 8(1):1–22, 2021.

[31] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

[33] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

[34] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[35] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2020.

[36] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

[37] David Thissen, Lynne Steinberg, and Daniel Kuang. Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of educational and behavioral statistics*, 27(1):77–83, 2002.

[38] Adam Yaari, Jan DeWitt, Henry Hu, Bennett Stankovits, Sue Felshin, Yevgeni Berzak, Helena Aparicio, Boris Katz, Ignacio Cases, and Andrei Barbu. The aligned multimodal movie treebank: An audio, video, dependency-parse treebank. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9531–9539. Association for Computational Linguistics, 2022. URL `https://aclanthology.org/2022.emnlp-main.648`.

[39] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

A Candidate DNN Models

Because they control for dataset and architecture (varying only the learning objective), comparisons amongst the variants of the SLIP models are our most empirically rigorous test of multimodality.

However, given that the SLIP models contain only one kind of multimodal - unimodal contrast (SLIP-SimCLR versus SLIP-combo's visual encoder), we added a number of uncontrolled model contrasts to assess the predictive power of unimodal and multimodal representations more generally. These models include VisualBERT (25) (a single-channel multimodal transformer that leverages cross-attention to integrate vision and language across all levels of representational hierarchy); SBERT (a unimodal masked language transformer for sentence embeddings); BEIT (a unimodal vision transformer trained via masked image reconstruction) (3); SimCSE (a unimodal language transformer trained via contrastive learning). These models provide a broader sample of multimodal and unimodal DNNs, while still maintaining some core similarities with the SLIP models (transformer backbones or contrastive learning.)

We assess both trained and randomly-initialized versions of these models first and foremost because, in most cases, the multimodality of these models is a function ONLY of their learning objective: This means, for example, that models like the SLIP models – which consist of architecturally encapsulated vision and language encoders – cannot, in the absence of training, be considered multimodal. Models like VisualBERT, on the other hand, may be considered multimodal even in the absence of training due to architectural inductive biases such as cross-modal attention-heads that integrate linguistic and visual inputs from the outset of processing. It's also worth noting that (especially in the case of language) randomly initialized networks are sometimes strong predictors of neural activity purely, it seems, as a function of architectural inductive biases (6; 36). This is a phenomenon that future work should investigate more thoroughly, but one we note preliminarily here.

B Neural Data Details

B.1 Event Structures

We parse our neural activity into individual event structures (language + single movie-frame combinations, which we call a text-image pair) by discretizing the movie stimulus, allowing us to feed inputs to our deep neural network models (which are not trained on movie data). We define event structures by the guiding feature used to select a particular text-image pair in the movie for analysis. So as not to unfairly prioritize one modality over the other, we design two different kinds of event structures: The first kind of event structure consists of word onset times, a language-aligned event. Word onsets have been used in prior work (18) and are commonly associated with language processing. For each word onset, we take the prior sentence context of the given word to add contextual information for the language models. We also take the closest frame after the word onset as the associated image input. The second kind of event structure consists of visual scene cuts (i.e. camera cuts). We extract the frames associated with a scene cut as proxy for visual processing given a shift in the pixel distribution between frames. We then take the closest sentence that occurred after the scene cut. (Note that by language-alignment or vision-alignment, here, we mean the anchoring of points in neural time-series to points in the movie).

We use these two kinds of event structures to create two datasets. Our language-aligned dataset consists of [context of a given word, closest frame pairs] with the associated neural activity as processed in Section 2. Our vision-aligned dataset consists of [scene cut frames, closest sentence to a scene cut frame] with similar processing on the neural activity. We analyze all results over the datasets individually and then compare results across the datasets to identify candidate electrodes for multimodal integration.

B.2 Notes on Stimulus Independence (Autoregression)

Converting neural activity measured in response to naturalistic movie-viewing to a dataset of nominally IID event-structures presents a particular challenge often explicitly avoided in experimental designs that leverage otherwise unrelated natural images or language prompts: that is, nonindependence in the form of autoregression. Movies (driven as they are by common visuolinguistic themes) contain inherently autoregressive structure that can lead to overfitting in parametrized predictive models designed to predict neural response patterns evoked by that structure. The parsing of our final event-structures into training, testing and validation splits was designed explicitly to assess for such overfitting. When creating the train-validation-test splits, we assign contiguous chunks of the movie to each split. In practice, and especially for movies with more linear narrative structure, we assumed this continguous splitting could provide at least a weak form of independence between sampled event-structures. While this by no means fully accounts for the non-independence of the stimulus set writ large, our results across the training, validation, and test splits suggests it does help to minimize potential overfitting. In future work, we hope to revisit our event-structure delineation and sampling, potentially leveraging movie-trained models like Salesforce's ALPRO (24) to select stimuli that are more distinct not just at the level of pixels or words, but in latent feature space.

C Neural Regressions

In this section, we detail our neural regression pipeline, which proceeds in 4 phases: feature extraction, dimensionality reduction (via sparse random projection), cross-validated ridge regression, and scoring.

C.1 Feature Extraction

This follows from approaches taken in Conwell et al. (11). We consider feature extraction to mean the extraction of a separate feature vector at *every layer* in a network – in other words, each distinct tensor operation module that progressively transforms model inputs into outputs. This means, for example, that we consider not only the outputs of each transformer attention head, but also of the individual key, query, value computations that produce them. If the layer is associated with a computation over visual features (e.g. BEIT, SimCLR) or multimodal features (e.g. VisualBERT), we flatten the tensor such that each layer represents any given input as a 1-dimensional feature vector. (Note: This flattening makes no assumptions about the separation of a given feature space into spatial and semantic components, and allows the subsequent regression to reweight all contributing components as relevant). If the layer is associated with a computation over language features (SBERT layer, CLIP language encoder layer), the output tensor will be consists of features over subwords. We average each representation associated with a subword to form word representations $\boldsymbol{P}$. The output tensor thus constitutes a dataset of n inputs (either images, sentences, or image-sentence pairs) as an array $\boldsymbol{F} \in \mathbb{R}^{n \times D}$ where D is the dimensions of the feature vector.

C.2 Sparse Random Projection

For certain flattened feature vectors from a particular, the dimensionality D is very large, and as such performing ridge regression on $\boldsymbol{F}$ is prohibitively expensive, with at best linear complexity with D, specifically $\mathcal{O}(n^2 D)$ (20). We use the Johnson-Lindenstrauss lemma (21; 13) to project $\boldsymbol{F}$ to a low dimensional representation $P \in \mathbb{R}^{n \times p}$ that preserves pairwise distances in $\boldsymbol{F}$ with errors bounded by a factor ϵ. If u and v are any two feature vectors from $\boldsymbol{F}$, and u_p and v_p are the low-dimensional projected vectors, then

$$(1-\epsilon)||u-v||^2 < ||u_p - v_p||^2 < (1+\epsilon)||u-v||^2 \tag{1}$$

Equation 1 holds provided that $p \geq \frac{4\ln(n)}{\epsilon^2/2 - \epsilon^3/3}$ (1). To find the mapping from $\boldsymbol{F}$ to $\boldsymbol{P}$, we used *sparse random projections* (SRPs) following Li et al. (26). The authors show a $\boldsymbol{P}$ satisfying Equation 1 can be found by $\boldsymbol{P} = \boldsymbol{F}\boldsymbol{R}$ where $\boldsymbol{R}$ is a sparse $n \times P$ matrix with i.i.d. elements shown below:

$$r_{ij} = \begin{cases} \sqrt{\frac{\sqrt{D}}{p}} & \text{with prob. } \frac{1}{2\sqrt{D}} \\ 0 & \text{with prob. } 1 - \frac{1}{\sqrt{D}} \\ -\sqrt{\frac{\sqrt{D}}{p}} & \text{with prob. } \frac{1}{2\sqrt{D}} \end{cases} \tag{2}$$

C.3 k-fold Ridge Regression

To determine how well vision and language networks predict activity in the brain, we ran regressions from representations extracted from a specific layer of either a multimodal or unimodal network to predict the average activity of the SEEG signals over a window of time for all electrodes of our 7 subjects. We detail the steps we took to run regressions per subject below.

We use ridge regression to predict the average activity, $\boldsymbol{y}$, at a given electrode and time point as constructed in Section 2, from their associated DNN features $\boldsymbol{P}$. Given the sequential nature of our data, we used a 5-fold cross-validation procedure. For each fold, we split our dataset of representations into a contiguous training set(80%), $\boldsymbol{P}_{\text{train}}$ and $\boldsymbol{y}_{\text{train}}$, a contiguous validation set (10%), $\boldsymbol{P}_{\text{valid}}$ and $\boldsymbol{y}_{\text{valid}}$, and contiguous testing set (10%), $\boldsymbol{P}_{\text{test}}$ and $\boldsymbol{y}_{\text{test}}$. Each split takes a contiguous chunk of event structures in order of their occurrence in the movie, and each fold changes the starting point of the training, validation, and testing set such that different contiguous chunks are assigned to a different set. We standardize the columns of $\boldsymbol{P}_{\text{train}}$ and $\boldsymbol{P}_{\text{valid}}$ to have mean 0 and a standard deviation of 1 and fit this standardization on $\boldsymbol{P}_{\text{test}}$. We fit the coefficients $\hat{\beta_i}$ of a regression model on the train dataset such that $\boldsymbol{y}_{\text{train}} = \boldsymbol{P}_{\text{train}}\hat{\beta_i} + \epsilon$ with minimal error $||\epsilon||$. Ridge regression penalizes large $||\hat{\beta}||$ proportional to a hyperparameter λ, which is useful in preventing overfitting when regressors are high-dimensional and highly correlated. Each $\hat{\beta}$ is calculated by the fixed ridge regression solution:

$$\hat{\beta} = ((\boldsymbol{P}_{\text{train}})^T \boldsymbol{P}_{\text{train}} + \lambda \boldsymbol{I}_d)^{-1} (\boldsymbol{P}_{\text{train}})^T \boldsymbol{y}_{\text{train}} \tag{3}$$

The coefficients $\hat{\beta}$ are then used to predict the held out data where:

$$\begin{aligned} \boldsymbol{y}_{\hat{\text{valid}}} &= \boldsymbol{P}_{\text{valid}}\beta \\ \boldsymbol{y}_{\hat{\text{test}}} &= \boldsymbol{P}_{\text{test}}\beta \end{aligned} \tag{4}$$

We use the *KFold* and *Ridge* functions from Pedregosa et al. (31). In this analysis we run the 5-fold regression per λ value, where λ was varied using a logarithmic grid search over 10^{-1} to 10^6. On each fold, we calculated a *score* for the prediction $\boldsymbol{y}_{\hat{\text{valid}}}$ and $\boldsymbol{y}_{\hat{\text{test}}}$ by computing the Pearson correlation coefficient. This score is averaged over the 5 folds to get final validation and test set scores. We choose the best λ value using the cross-validated scores and take the associated test scores with the λ value. We run this regression for all electrodes and time points simultaneously.

To analyze network performance over all layers, we select the best performing layer using the validation set. Specifically, per electrode, we average the validation correlation scores over time and take the layer with the max average score. We then take the associated test set correlation score as the overall score per model.

D Multimodality Test Details

Each of our multimodality tests in the main analysis is predicated on a bootstrapping procedure that probes the difference between two models in terms of their max decoding accuracy across the sliding time windows. We describe our bootstrapping analysis here and show a visualization in Figure 1(b). Consider the case for a single electrode and two models (model A and model B). An initial max over the test scores for all time windows will show one model to have a higher decoding accuracy than the other. To establish whether this difference is statistically significant (or just a product of noise across the sliding time windows), we perform 1000 bootstraps of the decoding scores across time windows, taking the difference between the higher and lower-ranking model each time. If the difference between the higher-ranking and lower-ranking model is greater than 0 in at least 950 / 1000 bootstraps (again equivalent to an alpha of 0.05), we register this difference as preliminarily significant (pending multiple comparison corrections).

We repeat this procedure for all electrodes for each of the model comparison tests we describe below. We then use FDR (Benjamini-Hochberg) (37) multiple comparisons corrections to adjust the p-value associated with each test on each electrode.

Each of the 4 tests we conduct are suggestive of multimodality, but each successive test provides additional evidence. After multiple comparison corrections, we tabulate the total number of electrodes that significantly pass each test as a proportion of the total number of assayed electrodes (N=1090). After aligning the location of the various electrodes to the regions provided by the Desikan-Killiany-Tourville atlas (22), we can further subdivide this proportion by the number of electrodes located in each region.