BrainTreebank-bench: Evaluating Foundation Models of Intracranial Brain Responses to Naturalistic Stimuli

Anonymous authors
Paper under double-blind review

Abstract

Foundation models have transformed fields from natural language processing to computer vision. Their great potential in neuroscience remains relatively untapped. We present BrainTreeBenchmark (BT-bench) as the next target for the advancement of foundation models of human intracranial brain signal. BT-bench contains 19 standardized decoding tasks (in the visual, auditory, language and multimodal categories), as well as defined train/test splits that evaluate performance within or across recording sessions, and within or across human subjects. BT-bench is based on the BrainTreebank dataset, a collection of intracranial neural data from patients undergoing clinical monitoring via implanted stereoelectroencephalography electrodes. The data were recorded while patients engaged in an ecological passive viewing paradigm, watching full-length Hollywood movies. We evaluate the performance of baseline decoding models on BT-bench and describe how BT-bench can enable tracking of information processing in the brain across tasks. Code to run BT-bench, as well as a public leaderboard website for community use, will be made available upon publication.

1 Introduction

Foundation models have driven rapid progress in domains such as natural language processing and computer vision. Given the high-dimensionality of neural signal and advances in the ability to obtain high-density brain recordings, there is immense potential for foundation models to transform neuroscience. This potential remains comparatively under-developed, however recent work points to

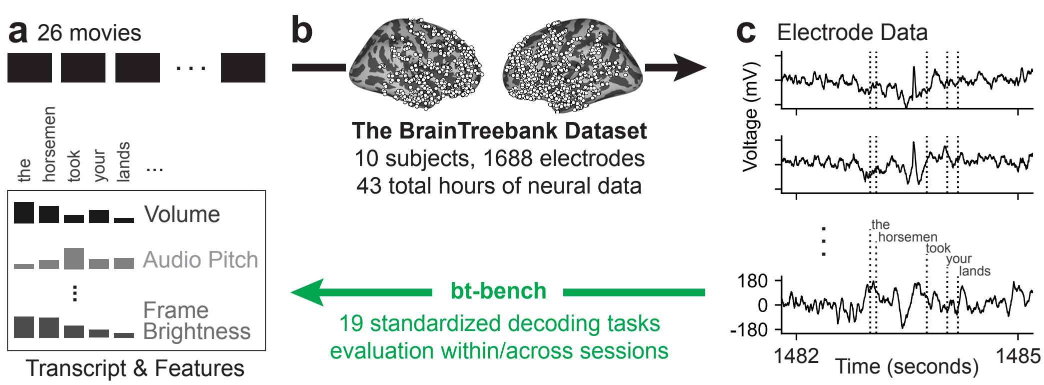

Figure 1: **Overview of BrainTreeBenchmark.** 26 movies (a) are watched by 10 epilepsy patients with stereoelectroencephalography electrodes implanted in various brain regions (b), and the local field potential from the implanted electrodes is available as part of the BrainTreebank dataset (c). BT-bench turns this dataset into an evaluation benchmark by segmenting the aligned data into various audio, language, and vision decoding tasks, such as, loudness and pitch of the audio, average pixel brightness, etc.

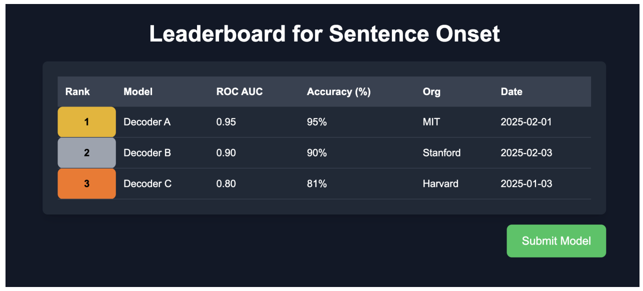

Figure 2: **The leaderboard for the task of classifying sentence onset.** The public webpage link will be made available upon publication.

a surge in large pretrained models based on neural activity: Neuroformer (Antoniades et al., 2024), BrainBERT (Wang et al., 2023), PopT (Chau et al., 2024), STNDT (Le & Shlizerman, 2022), NDT2 (Ye et al., 2023), MBrain (Cai et al., 2023), Brant (Zhang et al., 2023), MtM (Zhang et al., 2024), and POYO (Azabou et al., 2023).

There are a number of neural spiking activity datasets from non-human animals (for example, Perich et al. (2025); Churchland et al. (2024); Manley et al. (2024); IBL (2024)), as well as noninvasive recording technique datasets from humans, like fMRI (Wehbe et al., 2014; LeBel et al., 2023; Nastase et al., 2021; Li et al., 2022) and EEG (Zheng & Lu, 2015; Grootswagers et al., 2022; Bhattasali et al., 2020). Here we focus on intracranial human brain signal - specifically, stereoelectroencephalographic data (SEEG; for an overview, see Parvizi & Kastner (2018)). SEEG offers high temporal and spatial resolution that can reveal fundamental principles of cognition and language processing, yet no standard framework exists for benchmarking progress in modeling them.

BrainTreeBenchmark (BT-bench). We introduce BT-bench (Figure 1), a new suite of 19 standardized decoding tasks (Supplementary Table 2) derived from the BrainTreebank dataset, which contains intracranial recordings from multiple epilepsy patients watching annotated Hollywood films. Unlike smaller laboratory datasets, BT-bench leverages naturalistic stimuli and extensive annotations, providing a challenging test bed to evaluate modern representation learning methods.

Evaluations of neural decoders will be displayed on task-specific leaderboards (Figure 2) via our website. Machine learning engineers, neuroscientists, or anyone curious about the brain can follow the instructions, submit a model, and see how it compares to previous submissions. We establish well-defined train/test splits across sessions and subjects, allowing for rigorous within- and cross-subject generalization assessments (Table 1).

Train/Test Split	Description
SS-ST	Same Subject - Same Trial
SS-DT	Same Subject - Different Trial
DS-ST	Different Subject - Same Trial
DS-DT	Different Subject - Different Trial

Table 1: **Train/test split options for BT-bench.** The different splits allow for within- and cross-subject, as well as within- and cross- session generalization assessments.

The Brain Treebank Dataset. The Brain Treebank (Wang et al., 2024) is a large-scale dataset of intracranial electrophysiological recordings (stereoelectroencephalography; SEEG) collected while 10 human subjects (5 male, 5 female, ages 4–19; Supplementary Table 4) watched 26 total Hollywood movies (Supplementary Table 5). Electrode placements for each subject and their speech-selective responses are shown in Supplementary Figure 6. Spanning 43 hours of neural activity, the dataset aligns recorded brain signals with transcribed and manually corrected speech, word onsets, and universal dependency parses across the 223,068 words in 38,572 sentences. This dataset enables the systematic evaluation of computational models on multimodal neural decoding tasks.

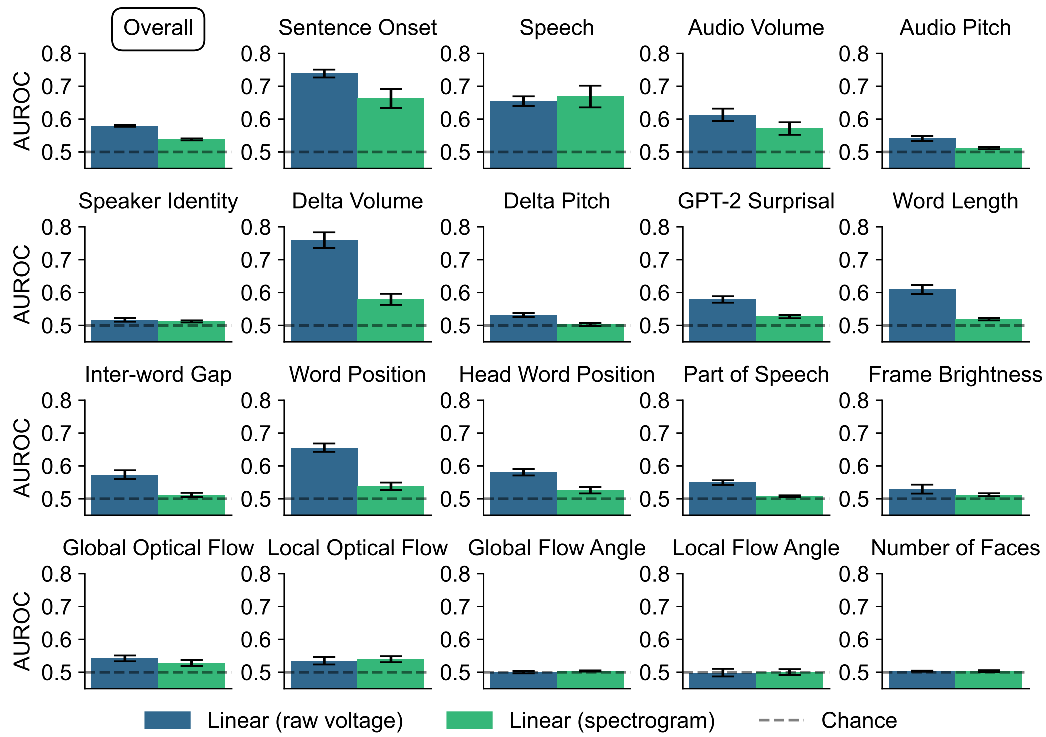

Figure 3: **Performance of baseline models on the 19 tasks of BT-bench.** Evaluation is done on the same subject, same trial (SS-ST), using 5-fold cross-validation. Normalized audio volume traces and the distribution of detected faces with corresponding word counts are shown in Supplementary Figures 5 and 7, respectively. The performance of two simple baseline models is shown: logistic regression (linear) either from raw voltage signal of all electrodes to the labels, or from the spectrogram of the signal to the labels. Neural data was cut to include one second following each word onset. In case of multi-class classification, AUROC was computed using a one-vs-all strategy and averaged together. Performance on different trials for the same subject were averaged together. Error bars denote s.e.m. across all subjects.

2 Evaluation

Comparison of basic decoding methods on BT-bench. We compare the performance of two simple baseline models—logistic regression applied to raw voltage signals and logistic regression applied to spectrogram features—across the 19 decoding tasks in BT-bench. Performance is evaluated using area under the receiver operating characteristic curve (AUROC), with chance-level performance ($ROC = 0.5$) included for reference.

Tracking of information processing in the brain across tasks. To investigate the time course of linguistic information processing in the brain, we aligned neural data to word onsets and split it into narrow time-bins (125ms width), training a separate linear decoder on each bin for multiple tasks. Decoding performance as a function of time shows a rise and fall after the word onset timestep, with the highest decoding performance achieved at a special point for every task (Figure 4). Interestingly, the beginning of a new sentence can be decoded even before the word onset, hinting at the predictive nature of processing.

3 Conclusion

We have presented the BrainTreeBenchmark, a suite of decoding tasks to measure the ability of foundation models to decode multimodal language processing in the brain. This benchmark has the

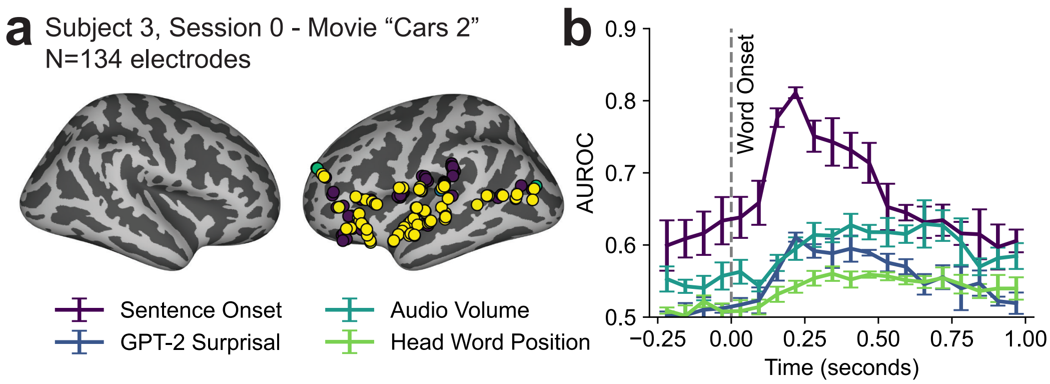

Figure 4: **BT-bench enables tracking of information processing in the brain across tasks.** (a) Decoding is run for all electrodes in a subject (subject 3; locations of electrodes plotted with the FDR-corrected p-value from 0 (yellow) to ≥ 0.1 (purple); see Supplementary Figure 6). (b) For this example trial, we trained a linear model across a sliding 125ms time window around word onset, and evaluated decoding performance as a function of time. Error bars show s.d. across the cross-validation runs.

potential to be used in two ways: (1) to probe the alignment between the internal representations of foundation models and the brain, as is done in Subramaniam et al. (2024), and (2) to track progress of fine-tuned foundation models to perform neural decoding tasks. This will drive improvements both in decoding ability and the ability to draw neuroscience conclusions from large scale data. As we have seen in other fields, this can also lead to a virtuous cycle in which neuroscientists are encouraged to share more datasets to the effort. By using our framework, any question about multimodal language processing in the brain can be posed as a machine learning task. Our framework is general enough to accommodate any future annotations, allowing for investigations of low-level language processing, such as part of speech, or high-level semantic processing such as thematic roles or language model embeddings.

We also seek, in near-term future work, to add to the library of tasks and datasets in BT-bench. As we continue to build out the benchmark, we will be able to study the question of how these tasks interact with each other. Each decoding task induces a map across the brain of when and where processing specific to that task is performed. By overlaying many of these maps, a functional picture of the brain can emerge of which language, vision, and audio features modulate activity in each region. We see this approach as a way of answering the long-standing neuroscience question: What is the underlying circuit basis of language processing in the brain?

REFERENCES

International brain lab. `https://internationalbrainlab.org`, 2024. Accessed: 2024-11-23.

Antonis Antoniades, Yiyi Yu, Joseph Canzano, William Wang, and Spencer LaVere Smith. Neuroformer: Multimodal and Multitask Generative Pretraining for Brain Data, March 2024.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael J. Mendelson, Blake Richards, Matthew G. Perich, Guillaume Lajoie, and Eva L. Dyer. A Unified, Scalable Framework for Neural Population Decoding, October 2023.

Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. The alice datasets: fMRI & EEG observations of natural language comprehension. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 120–125, Marseille, France, May 2020. European Language

Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.15/.

Donghong Cai, Junru Chen, Yang Yang, Teng Liu, and Yafeng Li. MBrain: A Multi-channel Self-Supervised Learning Framework for Brain Signals, June 2023.

Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population Transformer: Learning Population-level Representations of Neural Activity, October 2024.

Mark Churchland, John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy. Neural population dynamics during reaching. Data set, 2024. URL https://dandiarchive.org/dandiset/000070/draft.

Tijl Grootswagers, Iris Zhou, Austin K. Robinson, et al. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9:3, 2022. doi: 10.1038/s41597-021-01102-7. URL https://doi.org/10.1038/s41597-021-01102-7.

Trung Le and Eli Shlizerman. STNDT: Modeling Neural Population Activity with a Spatiotemporal Transformer, June 2022.

Alexandre LeBel, Laura Wagner, Siddharth Jain, et al. A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, 10:555, 2023. doi: 10.1038/s41597-023-02437-z. URL https://doi.org/10.1038/s41597-023-02437-z.

Jixing Li, Shohini Bhattasali, Shaolei Zhang, et al. Le petit prince multilingual naturalistic fmri corpus. *Scientific Data*, 9:530, 2022. doi: 10.1038/s41597-022-01625-7. URL https://doi.org/10.1038/s41597-022-01625-7.

Jason Manley, Sihao Lu, Kevin Barber, Jeffrey Demas, Hyewon Kim, David Meyer, Francisca Martínez Traub, and Alipasha Vaziri. Simultaneous, cortex-wide dynamics of up to 1 million neurons reveal unbounded scaling of dimensionality with neuron number. *Neuron*, 112(10):1694–1709.e5, 2024. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2024.02.011. URL https://www.sciencedirect.com/science/article/pii/S0896627324001211.

Samuel A. Nastase, Yung-Fang Liu, Harrison Hillman, et al. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8:250, 2021. doi: 10.1038/s41597-021-01033-3. URL https://doi.org/10.1038/s41597-021-01033-3.

Josef Parvizi and Sabine Kastner. Promises and limitations of human intracranial electroencephalography. *Nature Neuroscience*, 21(4):474–483, 2018. doi: 10.1038/s41593-018-0108-2. URL https://doi.org/10.1038/s41593-018-0108-2.

Matthew G. Perich, Lee E. Miller, Mehdi Azabou, and Eva L. Dyer. Long-term recordings of motor and premotor cortical spiking activity during reaching in monkeys. Data set, 2025. URL https://doi.org/10.48324/dandi.000688/0.250122.1735.

V Subramaniam, C Wang, A Barbu, G Kreiman, and B Katz. Revealing vision-language integration in the brain with multimodal networks. In *International Conference on Machine Learning*. International Conference on Machine Learning (ICML), 2024.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings, February 2023.

Christopher Wang, Adam Uri Yaari, Aaditya K Singh, Vighnesh Subramaniam, Dana Rosenfarb, Jan DeWitt, Pranav Misra, Joseph R. Madsen, Scellig Stone, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brain treebank: Large-scale intracranial recordings from naturalistic language stimuli, 2024. URL https://arxiv.org/abs/2411.08343.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS ONE*, 9(11):e112575, November 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0112575. URL `http://dx.plos.org/10.1371/journal.pone.0112575`.

Joel Ye, Jennifer L. Collinger, Leila Wehbe, and Robert Gaunt. Neural Data Transformer 2: Multicontext Pretraining for Neural Spiking Activity, September 2023.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation Model for Intracranial Neural Signal. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

Yizi Zhang, Yanchen Wang, Donato Jimenez-Beneto, Zixuan Wang, Mehdi Azabou, Blake Richards, Olivier Winter, International Brain Laboratory, Eva Dyer, Liam Paninski, and Cole Hurwitz. Towards a ”universal translator” for neural dynamics at single-cell, single-spike resolution, July 2024.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497.

A SUPPLEMENTARY INFORMATION

#	Feature	Description	Benchmark Task
1	frame_brightness *(visual)*	The mean brightness computed as the average HSV value over all pixels	Binary classification: low (percentiles 0%-25%) vs high (75%-100%)
2	global_flow *(visual)*	A camera motion proxy. The maximal average dense optical flow vector magnitude	Same as above
3	local_flow *(visual)*	A large displacement proxy. The maximal optical flow vector magnitude	Same as above
4	global_flow_angle *(visual)*	As 2, averaged over orientation (degrees) and selected by maximal magnitude	4-way classification: which of the cardinal directions is the closest
5	local_flow_angle *(visual)*	The orientation (degrees) of the largest local flow vector	Same as above
6	face_num *(visual)*	The maximum number of faces per frame during the word	3-way classification: 0, 1 or ≥ 2
7	volume *(auditory)*	Average root mean squared watts of the audio	Binary classification: low (0%-25%) vs high (75%-100%)
8	pitch *(auditory)*	Average pitch of the audio	Same as above
9	delta_volume *(auditory)*	The difference in average RMS of the 500ms windows pre- and post-word onset	Same as above
10	delta_pitch *(auditory)*	The difference in average pitch of the 500ms windows pre- and post-word onset	Same as above
11	speech *(language)*	Whether any speech is present in the given time interval	Binary classification
12	onset *(language)*	Whether a new sentence starts in the interval, or there is no speech at all	Binary classification
13	gpt2_surprisal *(language)*	Negative-log transformed GPT-2 word probability (given preceding 20s of language context)	Same as above
14	word_length *(language)*	Word length (ms)	Same as above
15	word_gap *(language)*	Difference between previous word offset and current word onset (ms)	Same as above
16	word_index *(language)*	The word index in its context sentence	4-way classification: 0, 1, 2, or ≥ 3
17	word_head_pos *(language)*	The relative position (left/right) of the word's dependency tree head	Binary classification
18	word_part_speech *(language)*	The word Universal Part-of-Speech (UPOS) tag	4-way classification: noun (0), verb (1), pronoun (2), or other (3)
19	speaker *(multimodal)*	The movie character that speaks the given word.	4-way classification: most frequent speaker (0), second (1), third (2), or other (3)

Table 2: **Extracted visual, auditory, and language features used to create the evaluations for BT-bench.** For all classification tasks, the classes were rebalanced. The difference between local and global flow is that global is the averaged optical flow, with the average being taken over all optical flow vectors on the screen, whereas local is the largest individual optical flow vector on the screen. The table is adapted from Chau et al. (2024).

Subj.	Age (yrs.)	# Electrodes	Movie	Recording time (hrs)	bt-bench testing
1	19	154	Thor: Ragnarok	1.83	
			Fantastic Mr. Fox	1.75	
			The Martian	0.5	
2	12	162	Venom	2.42	
			Spider-Man: Homecoming	2.42	
			Guardians of the Galaxy	2.5	x
			Guardians of the Galaxy 2	3	x
			Avengers: Infinity War	4.33	
			Black Panther	1.75	
			Aquaman	3.42	
3	18	134	Cars 2	1.92	x
			Lord of the Rings 1	2.67	
			Lord of the Rings 2 (extended edition)	3.92	
4	12	188	Incredibles	1.15	
			Shrek 3	1.68	
			Megamind	2.43	
5	6	156	Fantastic Mr. Fox	1.5	
6	9	164	Megamind	2.58	
			Toy Story	1.33	
			Coraline	1.83	
7	11	246	Cars 2	1.75	
			Megamind	1.77	
8	4.5	162	Sesame Street Episode	1.28	
9	16	106	Ant Man	2.28	
10	12	216	Cars 2	1.58	x
			Spider-Man: Far from Home	2.17	

Table 3: **Subject statistics** Subjects in the BrainTreebank dataset, and the trials used in the benchmark tasks. Table adapted from Wang et al. (2023). The second column shows the total number of electrodes. The average amount of recording data per subject is 4.3 (hrs).

Subj.	**Age**	**Sex**	**Movies**	**Time (h)**	**# Sent.**	**# Words**	**# Lemmas**	**# Elec.**	**# Probes**
1	19	M	7, 18, 19	5.6	4372	27424	4489	154	13
2	12	M	2, 3, 4, 8, 9, 17, 21	13.5	9870	57731	9164	162	47
3	18	F	5, 11, 12	7.5	5281	31596	4547	134	12
4	12	F	10, 13, 15	3.7	4056	23876	4017	188	15
5	6	M	7	1.35	1282	7908	1481	156	12
6	9	F	6, 13, 20	2.8	3789	20089	3349	164	12
7	11	F	5, 13	3.08	3523	19068	2828	246	18
8	4	M	14	0.94	860	3994	537	162	13
9	16	F	1	1.80	1558	9235	1480	106	12
10	12	M	5, 16	3.08	3981	22147	3004	216	17

Table 4: **All subjects language, electrodes and personal statistics.** Columns from left to right are the subject's ID and information (age and gender), the IDs of the movies they watched (corresponding to Supplementary Table 5), the cumulative movie time (hours), number of sentences, number of words (tokens) and number of unique lemmas (canonical word forms), as well as the number of probes the subject had and their corresponding number of electrodes. Table adapted from Wang et al. (2024).

#	Movie	Year	Length	Sent.	Words	Unique words	Nouns	Unique nouns	Verbs	Unique verbs
1	Antman	2015	7027	1558	9869	1944	1358	705	1545	580
2	Aquaman	2018	8601	1054	7233	1544	1069	520	1104	508
3	Avengers: Infinity War	2018	8961	1523	8529	1750	1083	607	1317	495
4	Black Panther	2018	8073	1254	7580	1606	1093	553	1209	508
5	Cars 2	2011	6377	2051	11407	2037	1572	724	1664	577
6	Coraline	2009	6036	997	5433	1232	784	409	805	348
7	Fantastic Mr. Fox	2009	5205	1282	8461	1864	1229	681	1227	484
8	Guardians of the Galaxy 1	2014	7251	1174	8295	1779	1096	603	1250	529
9	Guardians of the Galaxy 2	2017	8146	1290	9405	1824	1224	626	1370	532
10	Incredibles	2003	6926	1521	9430	1954	1226	652	1557	591
11	Lord of the Rings 1	2001	13699	1514	10566	1998	1473	679	1487	598
12	Lord of the Rings 2	2002	14131	1716	11041	2065	1588	743	1619	646
13	Megamind	2010	5735	1472	8891	1726	1172	602	1347	496
14	Sesame Street Ep. 3990	2016	3440	860	4220	787	717	231	706	217
15	Shrek the Third	2007	5568	1063	7226	1590	977	568	1071	422
16	Spiderman: Far From Home	2019	7764	1930	12189	1969	1459	668	1785	560
17	Spiderman: Homecoming	2017	8008	2196	12295	2066	1583	777	1808	572
18	The Martian	2015	9081	1570	11374	2192	1757	812	1677	622
19	Thor: Ragnarok	2017	7831	1583	9683	1789	1195	599	1419	548
20	Toy Story 1	1995	4863	1320	7216	1510	1019	548	1027	395
21	Venom	2018	6727	1379	7937	1513	897	507	1217	433

Table 5: **Language statistics for all movies.** Columns from left to right are the movie's ID, name, year of production, length (seconds), number of sentences, number of words (tokens), number of unique words (types), number of nouns, number of unique nouns, number of verbs and number of unique verbs. Table adapted from Wang et al. (2024).

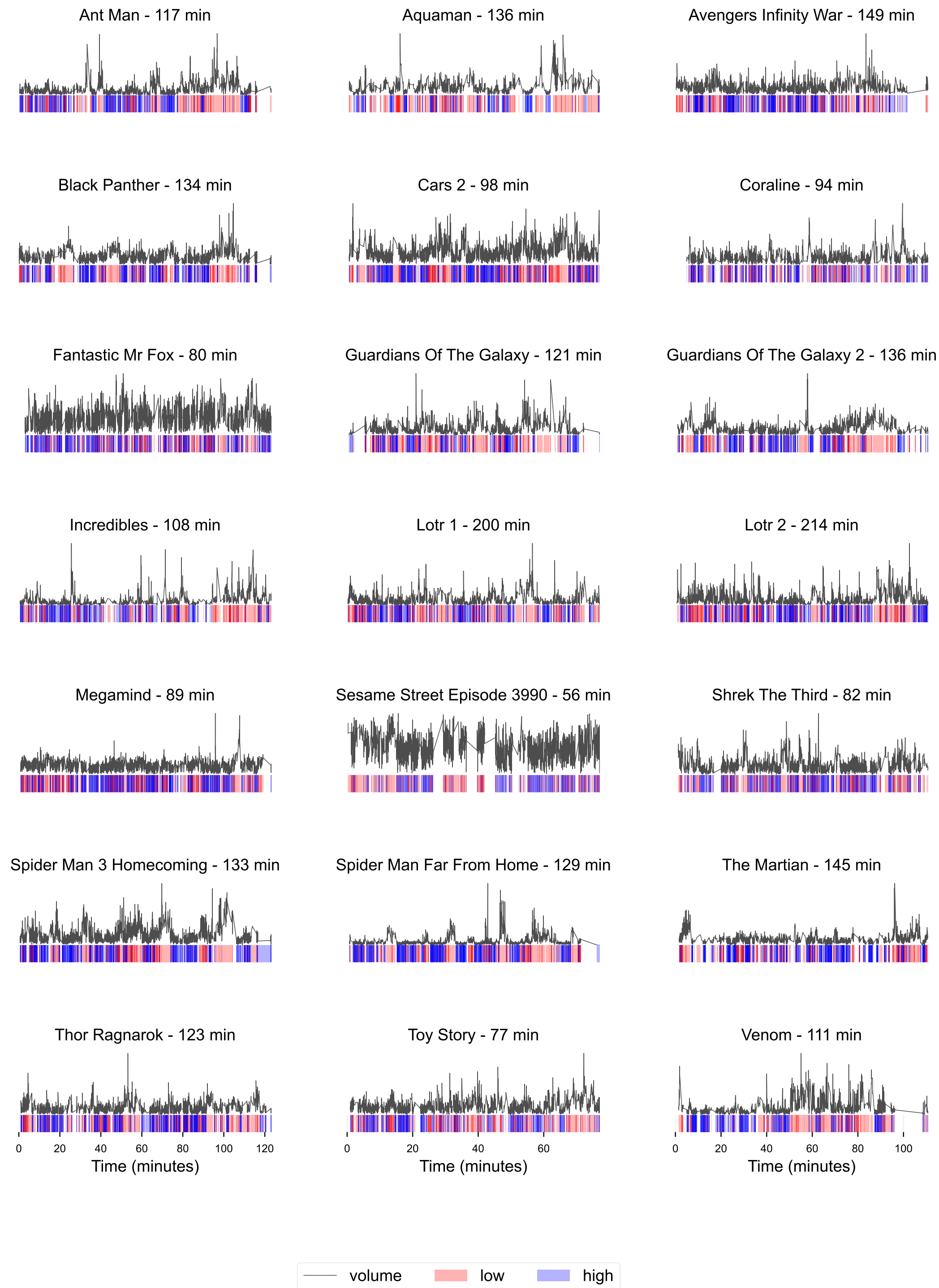

Figure 5: **Volume comparison across movies.** The black line shows the normalized audio volume over time for 18 feature-length films and one TV episode shown to subjects. Below each volume trace, colored bars indicate periods of relatively low (red) and high (blue) volume, defined as the bottom 25% and top 25% of volume values respectively.

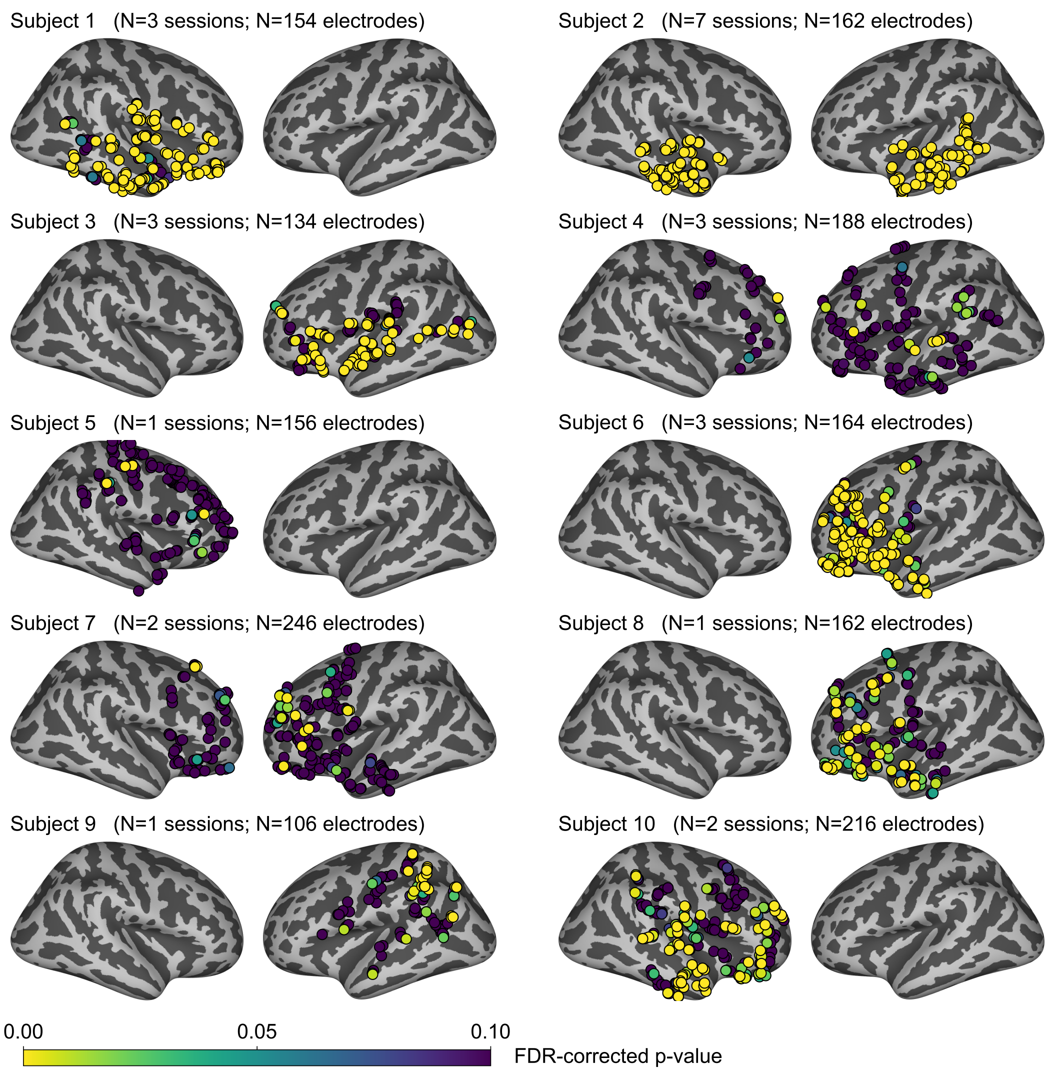

Figure 6: **Electrode locations and speech selectivity across subjects.** Brain reconstructions showing electrode placement and speech-selective responses for all 10 subjects. Each dot represents an electrode, colored by its FDR-corrected p-value from a speech vs. non-speech classification (color scale above, yellow indicating stronger selectivity). Left and right hemispheres are shown separately, with session counts and total electrodes noted. Speech selectivity was assessed by comparing high gamma power (70–300 Hz, dB) during the first 125 ms after word onset to non-speech intervals of equal duration. A two-sample t-test determined significance, with Benjamini-Hochberg correction applied for multiple comparisons.

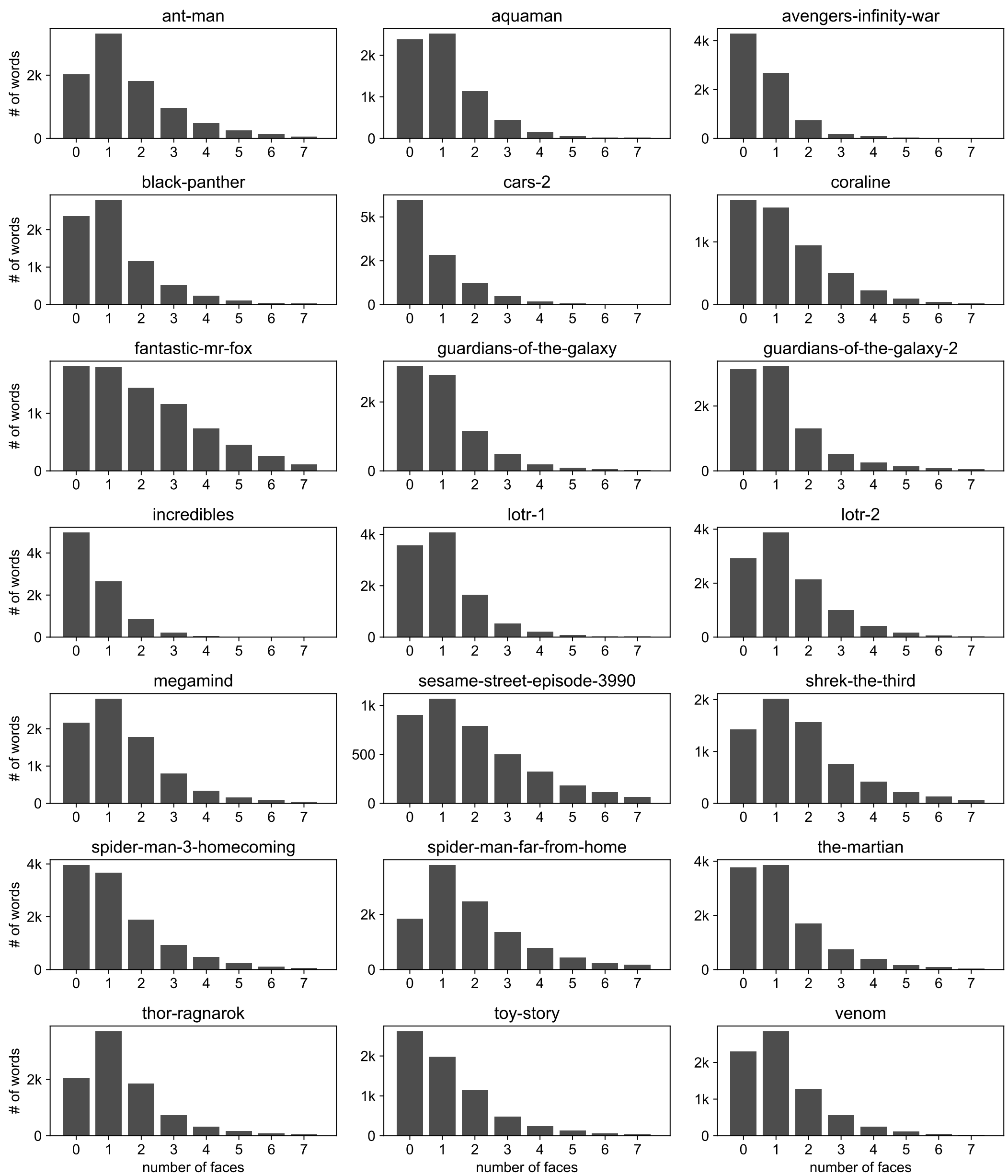

Figure 7: **Distribution of faces detected per frame across different movies.** Histograms show the number of words (y-axis) that occur during frames containing different numbers of faces (x-axis) for 18 feature-length films and one TV episode (Sesame Street)