Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey)

Anonymous authors
Paper under double-blind review

Abstract

Can we obtain insights about the brain using AI models? How is the information in deep learning models related to brain recordings? Can we improve AI models with the help of brain recordings? Such questions can be tackled by studying brain recordings like functional magnetic resonance imaging (fMRI). As a first step, the neuroscience community has contributed several large cognitive neuroscience datasets related to passive reading/listening/viewing of concept words, narratives, pictures, and movies. Encoding and decoding models using these datasets have also been proposed in the past two decades. These models serve as additional tools for basic cognitive science and neuroscience research. Encoding models aim at generating fMRI brain representations given a stimulus automatically. They have several practical applications in evaluating and diagnosing neurological conditions and thus may also help design therapies for brain damage. Decoding models solve the inverse problem of reconstructing the stimuli given the fMRI. They are useful for designing brain-machine or brain-computer interfaces. Inspired by the effectiveness of deep learning models for natural language processing, computer vision, and speech, several neural encoding and decoding models have been recently proposed. In this survey, we will first discuss popular representations of language, vision and speech stimuli, and present a summary of neuroscience datasets. Further, we will review popular deep learning based encoding and decoding architectures and note their benefits and limitations. Finally, we will conclude with a summary and discussion about future trends. Given the large amount of recently published work in the computational cognitive neuroscience (CCN) community, we believe that this survey enables an entry point for DNN researchers to diversify into CCN research.

1 Introduction

The central aim of neuroscience is to unravel how the brain represents information and processes it to carry out various tasks (visual, linguistic, auditory, etc.). Two important models related to how brain represents information are, how external stimuli are represented in the form of neural responses (*the encoding model*) and how stimuli are recovered or reconstructed from the neuronal responses (*the decoding model*). The recent progress in deep neural networks in processing visual, auditory, linguistic, and multimodal stimuli makes one wonder if we could investigate these computational models and shed light on how the brain solves these problems. Thus, deep neural networks (DNN) may offer a computational medium to capture brain activities unprecedented complexity and richness of, leading to accurate encoding and decoding solutions. Previous surveys (Cao et al., 2021; Karamolegkou et al., 2023) have primarily focused on brain encoding and decoding studies for language stimuli. But recent attempts in cognitive neuroscience have focused on naturalistic and multimodal stimuli using DNNs. Hence, this survey systematically summarizes the latest encoding and decoding efforts on (i) how DNNs have begun to explain the underlying information processing in the brain for naturalistic stimuli of various modalities, (ii) the ways in which DNN models may be improved using the brain data, and (iii) the exploration of the shared underlying characteristics of both the systems.

The survey aims to introduce the challenges in Computational Cognitive Neuroscience (CCN) to AI researchers familiar with recent advances in deep neural networks (DNNs). A good section of the DNN community is interested in neuroscience and psycholinguistics. Therefore, in this survey, we do not delve

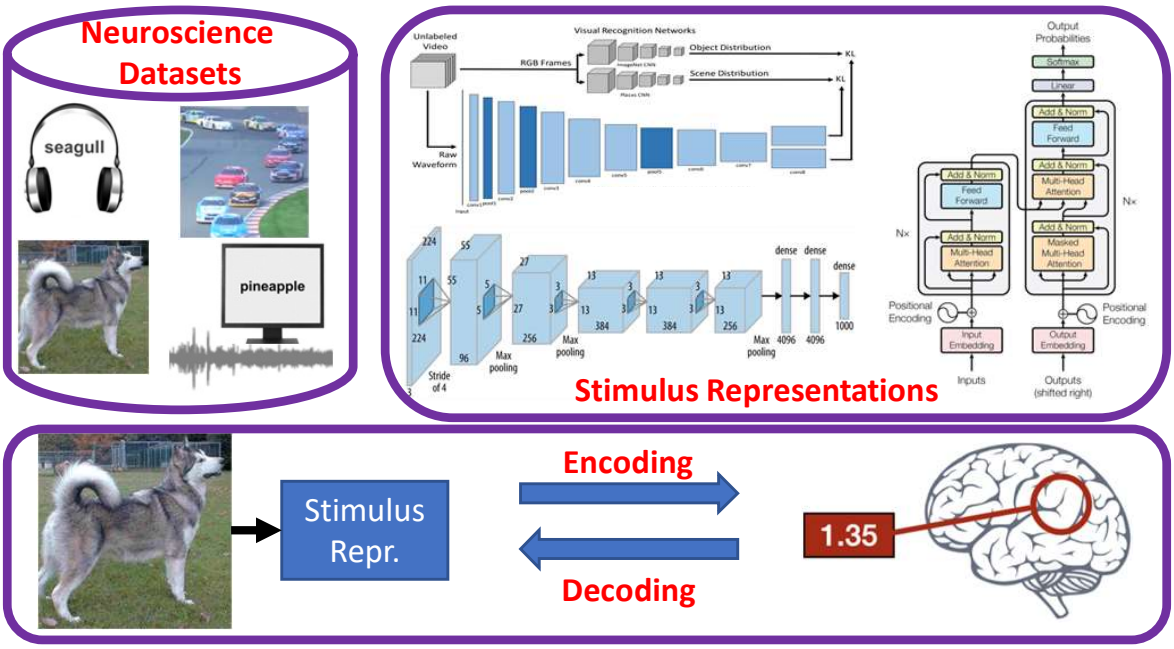


Figure 1: Brain Encoding and Decoding: Datasets & Stimulus Representations. The plot of the encoding-decoding framework was derived from the study by Ivanova et al. (2022).

into architectural details and the learning procedures for DNNs rather, highlight how the advances in DNNs are used to address CCN problems. This enables an entry point for DNN researchers to diversify into CCN research.

Specifically, DNN researchers have the following advantages:

1. Interpreting DNN models and evaluating their capabilities using naturalistic brain datasets.
2. Using open-source brain datasets as evaluation benchmarks to improve DNN model capabilities further.
3. Training DNN models by incorporating brain recordings.
4. Developing better brain-computer interface (BCI) capabilities to decode brain patterns using advanced DNN models.
5. Clear exposition of various open source ecological stimuli datasets available and a curated GitHub repository for quick start of a study.
6. An accessible taxonomy of models and approaches.
7. A collection of open research problems in this fast-breaking research domain.

Brain encoding and decoding. Two main tools studied in cognitive neuroscience are brain encoding and brain decoding, as shown in Figure 1. Encoding is the process of learning the mapping e from the stimuli S to the neural activation F. The mapping can be learned using features engineering or deep neural networks. On the other hand, decoding constitutes learning mapping d, which predicts stimuli S back from the brain activation F. However, in most cases, brain decoding aims to predict a stimulus representation R rather than reconstructing S. In both cases, the first step is to learn a semantic representation R of the stimuli S at the train time. Next, a regression function $e : R \rightarrow F$ is trained for encoding. For decoding, a function $d : F \rightarrow R$ is trained. These functions d and e can then be used at test time to process new stimuli and brain activations, respectively. Ridge regression is the most popular choice for the functions d and e.

To study the brain response to various modalities of stimuli, neuroscience researchers have curated several datasets. These datasets consist of stimuli and corresponding brain activity while participants were involved in interactions with the stimuli and optionally performing tasks such as language comprehension, visual and

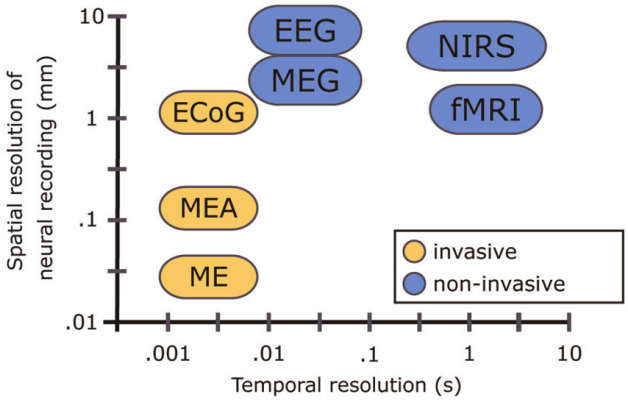


Figure 2: Overview of different brain–machine interfacing methods and their spatial and temporal resolution. Methods included: electroencephalography (EEG), magnetoencephalography (MEG), near-infrared spectroscopy (NIRS), functional magnetic resonance imaging (fMRI), electrocorticography (ECoG), microelectrode array (MEA) recordings and single microelectrode (ME) recordings. Adapted from (Van Gerven et al., 2009).

auditory processing, etc. Next, we discuss various techniques for obtaining the brain recordings and methods for representing stimuli.

Techniques for recording brain activations. Popular techniques for recording brain activations can be broadly classified into invasive and non-invasive techniques, as shown in Figure 2. Invasive techniques include single Micro-Electrode (ME), Micro-Electrode array (MEA), and Electro-Corticography (ECoG). The non-invasive recording techniques include functional magnetic resonance imaging (fMRI), Magneto-encephalography (MEG), Electro-encephalography (EEG) and Near-Infrared Spectroscopy (NIRS). Apart from the dimension of invasiveness, these techniques differ in their spatial resolution of neural recording and temporal resolution. fMRI recording enables data acquisition at high spatial but low temporal resolution. Hence, they are suitable for examining which parts of the brain handle critical functions. A typical whole brain fMRI acquisition takes 1-4 seconds to complete a scan. This is far slower than the speed at which humans can process language. On the other hand, both MEG and EEG have high temporal but low spatial resolution. They can preserve rich syntactic information (Hale et al., 2018) but cannot be used for source analysis. fNIRS offers a compromise option. The time resolution is better than fMRI, and spatial resolution is better than EEG. However, this spatial and temporal resolution balance may not compensate for the loss in both and its restriction in terms of only recording cortical activity but not from nuclei that are deeper in the brain, such as the basal ganglia, amygdala, hippocampus, etc. Further details on curation of brain recordings to specific brain regions are discussed in Section 4.

Stimulus representations. Neuroscience datasets contain stimuli across various modalities, including text, visual, audio, video, and other multimodal forms. Representations differ based on the modality. We briefly discuss extracting of stimulus representations from DNN models according to the following criteria: (1) Traditional and advanced models for text-based stimulus representations. (2) Image-based representations from deep vision models. (3) Extraction of low-level speech to Transformer-based speech-based auditory representations. (4) Finally, for multimodal stimulus representations, we explore early fusion and late fusion deep learning methods. Early fusion methods combine information across modalities at the initial processing stages, whereas late fusion combines it only at the end. Further details on different stimulus representation methods are discussed in Section 2.

Naturalistic neuroscience datasets. Several neuroscience datasets have been proposed across modalities (see Figure 3). These datasets differ in terms of the following criteria: (1) Method for recording activations: fMRI, EEG, MEG, etc. (2) Repetition time (TR), i.e. the sampling rate. (3) Characteristics of fixation points: location, color, shape. (4) Form of stimuli presentation: text, video, audio, images, or other multimodality. (5) Task that participant performs during recording sessions: question answering, property

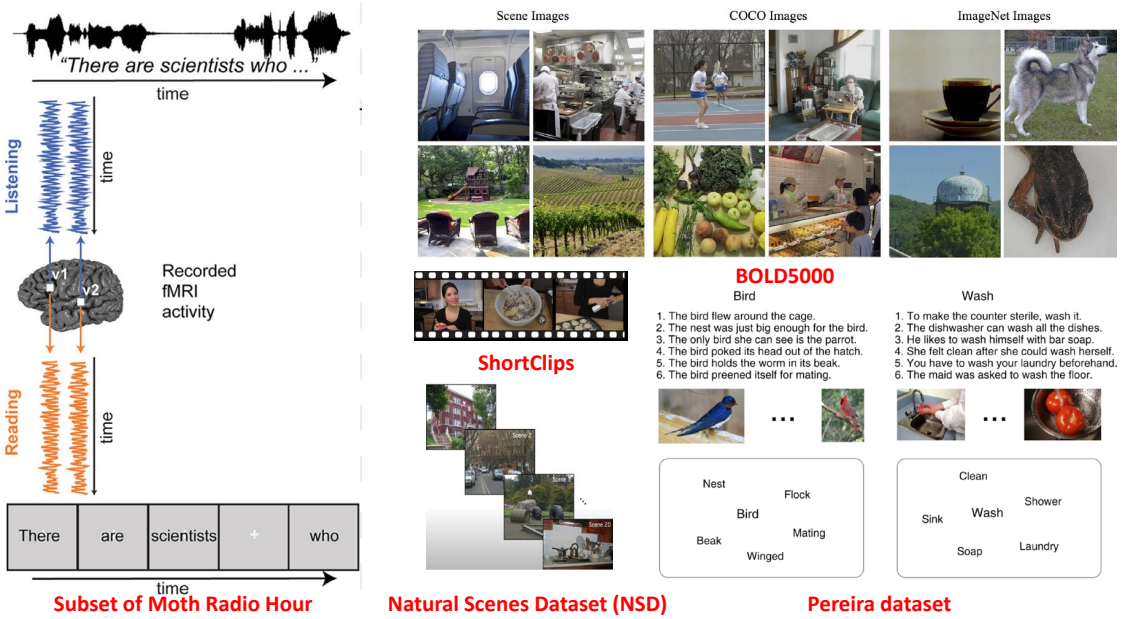


Figure 3: Representative Samples of Naturalistic Brain Datasets: (Left) Brain activity recorded when subjects are reading and listening to the same narrative (Deniz et al., 2019), and (Right) example naturalistic stimuli from various public repositories: BOLD5000 (Chang et al., 2019), ShortClips (Huth et al., 2022), Natural Scenes Dataset (NSD) (Allen et al., 2022) and Pereira dataset (Pereira et al., 2018).

generation, rating quality, etc. (6) Time given to participants for the task, e.g., 1 minute to list given object properties. (7) Demography of participants: males or females, sighted or blind, etc. (8) Number of times the response to stimuli was recorded. (9) Natural language associated with the stimuli. We discuss details of proposed datasets in Section 3.

Evaluation of brain encoding and decoding methods. 2V2 accuracy and Pearson Correlation are two popularly used metrics for the evaluation of brain encoding models. On the other hand, brain decoding models are evaluated using metrics such as pairwise accuracy, rank accuracy, R^2 score, and mean squared error. We discuss the detailed definitions of these metrics in Section 5.

Interpreting brain recordings through the robustness of DNN model representations. To interpret the stimulus representations obtained from DNN models and examine their impact on brain alignment, prior studies have proposed three different methods: variance partitioning (de Heer et al., 2017), the residual approach (Toneva et al., 2022a; Oota et al., 2024b), an indirect approach (Schrimpf et al., 2021; Goldstein et al., 2022), and the stacked regression approach (Lin et al., 2023). We discuss the details of each method in Section 6.3.

Computational Cognitive Neuroscience (CCN) research goals. CCN researchers have primarily focused on two main areas (Doerig et al., 2023).

1. Improving predictive accuracy. In this area, the work is around the following questions.
   - Compare feature sets: Which feature set provides the most faithful reflection of the neural representational space?
   - Test feature decodability: "Does neural data Y contain information about features X?"
   - Build accurate models of brain data: The aim is to enable the simulation of neuroscience experiments.
2. Interpretability. In this area, the work is around the following questions.
   - Examine individual features: Which contributes most to neural activity?

- Test correspondences between representational spaces: "CNNs vs ventral visual stream" or "Two text representations".
- Interpret feature sets: Do features X, generated by a known process, accurately describe the space of neural responses Y? Do voxels respond to a single feature or exhibit mixed selectivity?
- How does the mapping relate to other brain function models or theories?

We discuss some of these questions in Sections 6 and 7.

Brain encoding literature (Mitchell et al., 2008; Wehbe et al., 2014; Huth et al., 2016) has focused on studying several important aspects: (1) Which models lead to better predictive accuracy across modalities? (Toneva & Wehbe, 2019; Deniz et al., 2019; Schrimpf et al., 2021) (2) How can we disentangle the contributions of syntax and semantics from language model representations to the alignment between brain recordings and language models? (Lopopolo et al., 2017; Reddy & Wehbe, 2021) (3) Why do some representations lead to better brain predictions? How are deep learning models and brains aligned in terms of their information processing pipelines? (Merlin & Toneva, 2022; Aw & Toneva, 2023) (4) Does joint encoding of task and stimulus representation help? (Oota et al., 2024b). We discuss these details of encoding methods in Section 6.

Brain decoding models aim to understand what a subject is thinking, seeing, and perceiving by analyzing neural recordings. Over the past decades, the brain-computer interface (BCI) has made significant progress in decoding stimuli (language/images/speech) from the brain using non-invasive recordings. Like brain encoding literature, decoding literature focuses on studying a few important aspects: (1) In the context of language, how we compose the linguistic meaning from different stimuli such as text, images, videos, or speech by analyzing the evoked brain activity (Pereira et al., 2016; 2018). (2) Given brain activations corresponding to visual stimuli, how accurately can we decode a sentence representing the visual stimuli? (Nishimoto et al., 2011; Beliy et al., 2019) (3) How can we decode natural speech processing from non-invasive brain recordings using a single architecture and a data-driven approach? (Défossez et al., 2023) (4) How accurately can we reconstruct perceived natural images or decode their semantic contents from non-invasive recording data using popular deep learning models? (Takagi & Nishimoto, 2022). We discuss these details of decoding methods in Section 7.

2 Stimulus Representations

In this section, we discuss types of stimulus representations proposed in the literature across different modalities: text, visual, audio, video, and other multimodal stimuli.

Text stimulus representations. Older methods for text-based stimuli representation include text corpus co-occurrence counts (Mitchell et al., 2008; Pereira et al., 2013; Huth et al., 2016), topic models (Pereira et al., 2013), syntactic features and discourse features (Wehbe et al., 2014). In recent times, for text-based stimuli, both semantic models and experiential attribute models have been explored. Semantic representation models include word embedding methods (Pereira et al., 2018; Wang et al., 2020; Pereira et al., 2016; Toneva & Wehbe, 2019; Anderson et al., 2017a; Oota et al., 2018), sentence representation models (Sun et al., 2020; 2019; Toneva & Wehbe, 2019), RNNs (Jain & Huth, 2018; Oota et al., 2019) and Transformer methods (Gauthier & Levy, 2019; Toneva & Wehbe, 2019; Schwartz et al., 2019; Schrimpf et al., 2021; Antonello et al., 2021; Oota et al., 2022b; Aw & Toneva, 2023). Popular word embedding methods include textual (i.e., Word2Vec (Mikolov et al., 2013), fastText (Bojanowski et al., 2017), and GloVe (Pennington et al., 2014)), linguistic (i.e., dependency), conceptual (i.e., RWSGwn (Goikoetxea et al., 2015) and ConceptNet (Speer et al., 2017)), contextual (i.e., ELMo (Peters et al., 2018)). Popular sentence embedding models include average, max, concat of avg and max, SIF (Arora et al., 2017), SkipThoughts (Kiros et al., 2015), GenSen (Subramanian et al., 2018), InferSent (Conneau et al., 2017), ELMo, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), USE (Cer et al., 2018), QuickThoughts (Logeswaran & Lee, 2018) and GPT-2 (Radford et al., 2019). Transformer-based methods include pretrained BERT with various NLU tasks, finetuned BERT, Transformer-XL (Dai et al., 2019), GPT-2, BART (Lewis et al., 2020), BigBird (Zaheer et al., 2020), Longformer (Beltagy et al., 2020), and LongT5 (Guo et al., 2022). Experiential attribute models represent words in terms of human ratings of their degree of association with different attributes of

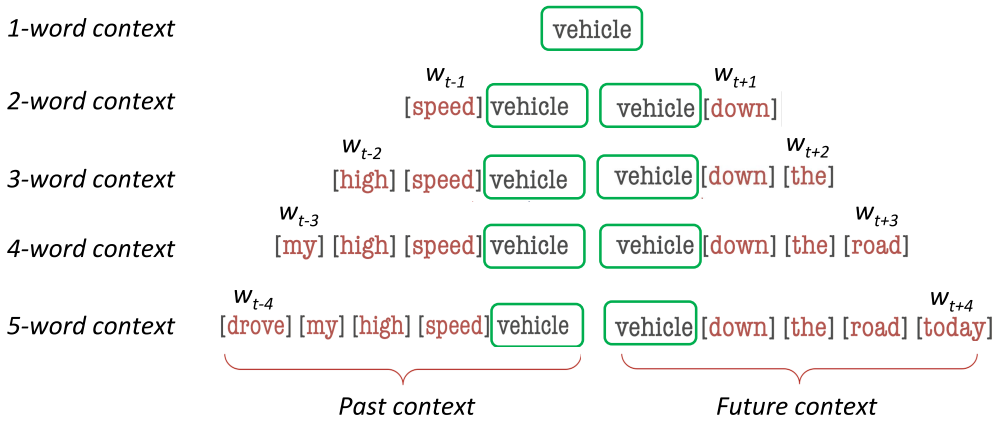


Figure 4: *Context representation of several word orders:* Past/Future context is constructed by considering words preceding/succeeding the current word (see *Past/Future context* illustrated for the current word *vehicle* for various orders).

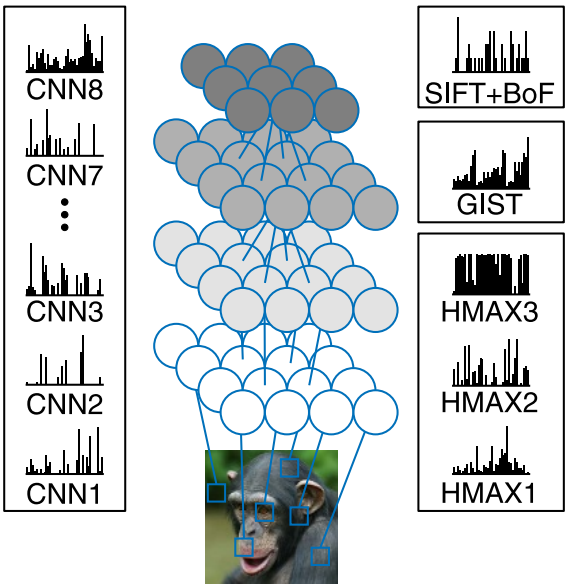


Figure 5: *Extraction of image representations:* Prior research has explored the impact of various layer-wise image representations from CNN models (Yamins et al., 2014; Horikawa & Kamitani, 2017), for both brain encoding and decoding models. The plot of the image feature extraction was derived from the study by (Horikawa & Kamitani, 2017).

experience, typically on a scale of 0-6 (Anderson et al., 2019; 2020; Berezutskaya et al., 2020; Just et al., 2010; Anderson et al., 2017b) or binary (Handjaras et al., 2016; Wang et al., 2017).

In the practice of employing word embeddings, encoding studies often utilize the average word representations within a given context or derive complete sentence representations through sentence embedding models. More recently, brain encoding research has shifted towards the use of contextualized word representations, examining how the amount of context affects the brain predictivity (Jain & Huth, 2018; Toneva & Wehbe, 2019). To obtain these contextualized word representations, Figure 4 illustrates how Past/Future context is constructed by considering words preceding/succeeding the current word. Given the constrained context length, each word is successively input to the network with at most C previous tokens. For instance, given a story of M words and considering the context length of 20, while the third word's vector is computed by

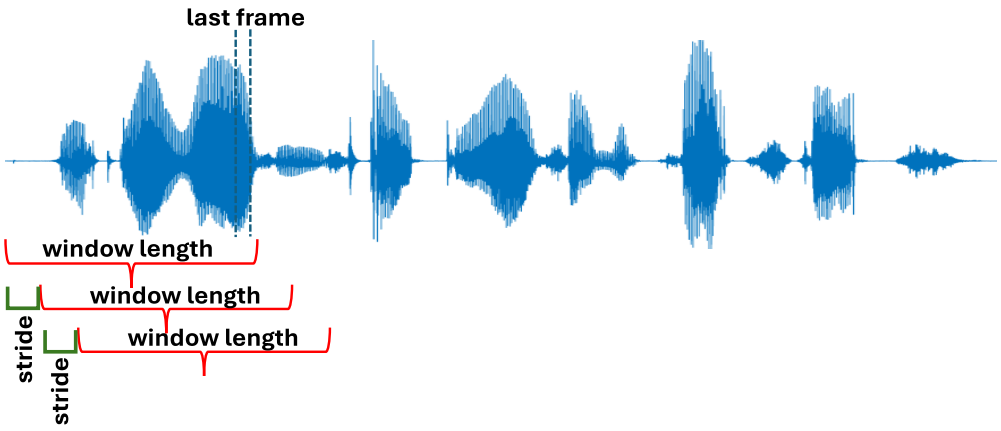


Figure 6: *Extraction of contextualized speech representations:* Representation of the last frame within each window allows for the capture of temporal dynamics and contextual nuances in the speech signal. The length of the time window is typically varied from 16, 32, to 64 secs, with strides ranging from 10 to 100 milliseconds.

inputting the network with (w_1, w_2, w_3), the last word's vectors w_M is computed by inputting the network with (w_{M-20}, ..., w_M).

Visual stimulus representations. For visual stimuli, older methods used visual field filter bank (Thirion et al., 2006; Nishimoto et al., 2011) and Gabor wavelet pyramid (Kay et al., 2008; Naselaris et al., 2009). As shown in Figure 5, recent methods use models like CNNs (Du et al., 2020; Beliy et al., 2019; Anderson et al., 2017a; Yamins et al., 2014; Nishida et al., 2020) and concept recognition models (Anderson et al., 2020).

Audio stimuli representations. For audio stimuli, phoneme rate and presence of phonemes have been leveraged (Huth et al., 2016). Further, low-level speech features like filter banks (FBank), Mel Spectrogram, and MFCC from raw audio files, phonological features, articulation, and power spectrum (PowSpec) feature vectors were used in (Deniz et al., 2019). Recently, Nishida et al. (2020) used features from an audio deep learning model called SoundNet for audio stimuli representation. To extract representations from Transformer-based speech models such as Wav2Vec2.0, HuBERT and Whisper, Vaidya et al. (2022); Antonello et al. (2024); Oota et al. (2024a) varied the length of the time windows from 16, 32, to 64 seconds, with strides ranging from 10 to 100 milliseconds, as illustrated in Figure 6. Moreover, these studies utilized an autoregressive approach to derive speech representations. This method involves considering the representations of the last frame within each window, allowing for the capture of temporal dynamics and contextual nuances in speech.

Multimodal stimulus representations. To jointly model the information from multimodal stimuli, recently, various multimodal representations have been used. These include processing videos using audio+image representations like VGG (Simonyan & Zisserman, 2015) and SoundNet (Aytar et al., 2016) in (Nishida et al., 2020) or using image+text combination models like GloVe+VGG and ELMo+VGG in (Wang et al., 2020). Recently, the usage of multimodal text+vision models like Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021), Learning Cross-Modality Encoder Representations from Transformers (LXMERT) (Tan & Bansal, 2019), and VisualBERT (Li et al., 2020) was proposed in (Oota et al., 2022e).

3 Naturalistic Neuroscience Datasets

In this section, we discuss the popular text, visual, audio, video, and other multimodal neuroscience datasets that have been proposed in the literature. Tables 1 and 2 show a detailed overview of brain recording type,

Table 1: Naturalistic Neuroscience Datasets (Text and Audio). Publicly available datasets are linked to their sources in the Dataset column. In this table, |S| represents the number of participants in each dataset.

	Dataset	Authors	Type	Lang.	Stimulus	|S|	Task
Text	Harry Potter	(Wehbe et al., 2014)	fMRI, MEG	English	Reading Chapter 9 of Harry Potter and the Sorcerer's Stone	9	Story understanding
	-	(Handjaras et al., 2016)	fMRI	Italian	Verbal, pictorial or auditory presentation of 40 concrete nouns, four times	20	Property generation
	-	(Anderson et al., 2017a)	fMRI	Italian	Reading 70 concrete and abstract nouns from law/music, five times	7	Imagine a situation with noun
	ZuCo	(Hollenstein et al., 2018)	EEG	English	Reading 1107 sentences with 21,629 words from movie reviews	12	Rate movie quality
	240 Sentences with Content Words	(Anderson et al., 2019)	fMRI	English	Reading 240 active voice sentences describing everyday situations	14	Passive reading
	BCCWJ-EEG	(Oseki & Asahara, 2020)	EEG	Japanese	Reading 20 newspaper articles for ∼30-40 minutes	40	Passive reading
	Subset Moth Radio Hour	(Deniz et al., 2019)	fMRI	English	Reading 11 stories	9	Passive reading and listening
Audio	-	(Handjaras et al., 2016)	fMRI	Italian	Verbal, pictorial or auditory presentation of 40 concrete nouns, 4 times	20	Property generation
	The Moth Radio Hour	(Huth et al., 2016)	fMRI	English	Listening eleven 10-minute stories	7	Passive listening
	Narrative Brain Dataset	(Lopopolo et al., 2018)	fMRI	Dutch	Spoken presentation of short excerpts of three stories	24	Passive listening
	Alice	(Brennan & Hale, 2019)	EEG	English	Listening Chapter one of Alice's Adventures in Wonderland (2,129 words in 84 sentences) as read by Kristen McQuillan	33	Question answering
	-	(Anderson et al., 2020)	fMRI	English	Listening one of 20 scenario names, 5 times	26	Imagine personal experiences
	Narratives	(Nastase et al., 2020)	fMRI	English	Listening 27 diverse naturalistic spoken stories. 891 functional scans	345	Passive listening
	Natural Stories	(Zhang et al., 2020)	fMRI	English	Listening Moth-Radio-Hour naturalistic spoken stories.	19	Passive listening
	The Little Prince	(Li et al., 2021)	fMRI	English, Chinese, French	Listening audiobook for about 100 minutes.	112	Passive listening
	MEG-MASC	(Gwilliams et al., 2023)	MEG	English	Listening two hours of naturalistic stories. 208 MEG sensors	27	Passive listening
	Music Genre	(Nakai et al., 2022)	fMRI	English	Listening 540 music pieces from 10 music genres	5	Passive listening
	SMN4Lang	(Wang et al., 2022b)	fMRI, MEG	Chinese	Listening 6 hours of naturalistic stories	12	Passive listening

language, stimulus, number of subjects (|S|), and the task across datasets of different modalities. Figure 3 shows examples from a few datasets. A sample of naturalistic datasets is available at this link [1].

Text datasets. These datasets are created by presenting words, sentences, passages, or chapters as stimuli. Some of the text datasets include Harry Potter Story (Wehbe et al., 2014), ZuCo EEG (Hollenstein et al., 2018) and datasets proposed in (Handjaras et al., 2016; Anderson et al., 2017a; 2019; Wehbe et al., 2014). In Handjaras et al. (2016), participants were asked to verbally enumerate in one minute the properties (features) that describe the entities the words refer to. There were four groups of participants: 5 sighted individuals were presented with a pictorial form of the nouns, 5 sighted individuals with a verbal-visual (i.e., written Italian words) form, 5 sighted individuals with a verbal auditory (i.e., spoken Italian words) form, and 5 congenitally blind with a verbal auditory form. Data proposed by Anderson et al. (2017a) contains 70 Italian words taken from seven taxonomic categories (abstract, attribute, communication, event/action, person/social role, location, object/tool) in the law and music domain. The word list contains concrete as well as abstract words. ZuCo dataset (Hollenstein et al., 2018) contains sentences for which EEG recordings were obtained for 3 tasks: normal reading of movie reviews, normal reading of Wikipedia sentences, and task-specific reading of Wikipedia sentences. For this dataset curation, sentences were presented to the

[1] https://neuroscout.org/datasets

Table 2: Naturalistic Neuroscience Datasets (Visual, Video, and Other Multimodal). Publicly available datasets are linked to their sources in the Dataset column. In this table, |S| represents the number of participants in each dataset.

	Dataset	Authors	Type	Lang.	Stimulus	|S|	Task
Visual	Inverse retinotopy	(Thirion et al., 2006)	fMRI	-	Viewing rotating wedges (8 times), expanding/contracting rings (8 times), rotating 36 Gabor filters (4 times), grid (36 times)	9	Passive viewing
	Vim-1	(Kay et al., 2008)	fMRI	-	Viewing sequences of 1870 natural photos	2	Passive viewing
	Generic Object Decoder	(Horikawa & Kamitani, 2017)	fMRI	-	Viewing 1,200 images from 150 object categories; 50 images from 50 object categories; imagery 10 times	5	Repetition detection
	BOLD5000	(Chang et al., 2019)	fMRI	-	Viewing 5254 images depicting real-world scenes	4	Passive viewing
	Algonauts	(Cichy et al., 2019)	fMRI, MEG	-	Viewing 92 silhouette object images and 118 images of objects on natural background	15	Passive viewing
	NSD	(Allen et al., 2022)	fMRI	-	Viewing 73000 natural scenes	8	Passive viewing
	THINGS	(Hebart et al., 2023)	fMRI, MEG	-	Viewing 31188 natural images	8	Oddball Detection
	NOD	(Gong et al., 2023)	fMRI	-	Viewing 57,120 natural images	30	Passive viewing
Video	BBC's Doctor Who	(Seeliger et al., 2019)	fMRI	English	Viewing spatiotemporal visual and auditory videos (30 episodes). 120.8 whole-brain volumes (∼23 h) of single-presentation data, and 1.2 volumes (11 min) of repeated narrative short episodes. 22 repetitions	1	Passive viewing
	Japanese Ads	(Nishida et al., 2020)	fMRI	Japanese	Viewing 368 web and 2452 TV Japanese ad movies (15-30s). 7200 train and 1200 test fMRIs for web; fMRIs from 420 ads.	52	Passive viewing
	Pippi Langkous	(Berezutskaya et al., 2020)	ECoG	Swedish, Dutch	Viewing 30 s excerpts of a feature film (in total, 6.5 min long), edited together for a coherent story	37	Passive viewing
	Algonauts	(Cichy et al., 2021)	fMRI	English	Viewing 1000 short video clips (3 sec each)	10	Passive viewing
	Natural Short Clips	(Huth et al., 2022)	fMRI	English	Watching natural short movie clips	5	Passive viewing
	Natural Short Clips	(Lahner et al., 2023)	fMRI	English	Watching 1102 natural short video clips	10	Passive viewing
	NNDb	(Aliko et al., 2020)	fMRI	English	Watching 10 full-length movies	84	Passive viewing
	NATVIEW_EEGFMRI	(Telesford et al., 2023)	fMRI, EEG	English	Watching 5 short-length movies	22	Passive viewing
	Mind captioning	(Horikawa, 2024)	fMRI	English	Watching total of 2,196 videos	6	Passive viewing
Other	60 Concrete Nouns	(Mitchell et al., 2008)	fMRI	English	Viewing 60 different word-picture pairs from 12 categories, 6 times each	9	Passive viewing
	-	(Sudre et al., 2012)	MEG	English	Reading 60 concrete nouns along with line drawings. 20 questions per noun lead to 1200 examples.	9	Question answering
	-	(Zinszer et al., 2018)	fNIRS	English	8 concrete nouns (audiovisual word and picture stimuli): bunny, bear, kitty, dog, mouth, foot, hand, and nose; 12 times repeated.	24	Passive viewing and listening
	Pereira	(Pereira et al., 2018)	fMRI	English	Viewing 180 Words with Picture, Sentences, word clouds; reading 96 text passages; 72 passages. 3 times repeated.	16	Passive viewing and reading
	-	(Cao et al., 2021)	fNIRS	Chinese	Viewing and listening 50 concrete nouns from 10 semantic categories.	7	Passive viewing and listening
	Neuromod	(Boyle et al., 2020)	fMRI	English	Watching TV series and movies (Friends, Movie10)	6	Passive viewing and listening
	Multimodal fMRI	(Jung et al., 2024)	fMRI	English	Watching movies, dynamic faces task	101	Passive viewing and listening

subjects in a naturalistic reading scenario. A complete sentence is presented on the screen. Subjects read each sentence at their own speed, i.e., the reader determines for how long each word is fixated and which word to fixate next.

Visual datasets. Older visual datasets were based on binary visual patterns (Thirion et al., 2006). Recent datasets contain natural images. Examples include Vim-1 (Kay et al., 2008), BOLD5000 (Chang et al., 2019), Algonauts (Cichy et al., 2019), NSD (Allen et al., 2022), Things-data (Hebart et al., 2023), NOD (Gong et al., 2023), and the dataset proposed in (Horikawa & Kamitani, 2017). BOLD5000 includes ∼20 hours of MRI scans per each of the four participants. 4,916 unique images were used as stimuli from 3 image sources. Algonauts contains two sets of training data, each consisting of an image set and brain activity in RDM format (for fMRI and MEG). Training set 1 has 92 silhouette object images, and training set 2 has 118 object images with natural backgrounds. Testing data consists of 78 images of objects on natural backgrounds. Most of the visual datasets involve passive viewing, but the dataset in (Horikawa & Kamitani, 2017) involved the participant doing the one-back repetition detection task.

Audio datasets. Most of the proposed audio datasets are in English (Huth et al., 2016; Brennan & Hale, 2019; Anderson et al., 2020; Nastase et al., 2020), while there is one (Handjaras et al., 2016) on Italian, and another one (Li et al., 2021) in Chinese and French. The participants were involved in a variety of tasks while their brain activations were measured: Property generation (Handjaras et al., 2016), passive listening (Huth et al., 2016; Nastase et al., 2020), question answering (Brennan & Hale, 2019) and imagining themselves personally experiencing common scenarios (Anderson et al., 2020). In the last one, participants underwent fMRI as they reimagined the scenarios (e.g., resting, reading, writing, bathing, etc.) when prompted by standardized cues. Narratives (Nastase et al., 2020) used 27 different stories as stimuli. Across subjects, it is 6.4 days worth of recordings.

Video datasets. Recently, video neuroscience datasets have also been proposed. These include BBC's Doctor Who (Seeliger et al., 2019), Japanese Ads (Nishida et al., 2020), Pippi Langkous (Anderson et al., 2020) and Algonauts (Cichy et al., 2021). Japanese Ads data contains data for two sets of movies provided by NTT DATA Corp: web and TV ads. There are also four types of cognitive labels associated with the movie datasets: scene descriptions, impression ratings, ad effectiveness indices, and ad preference votes. Algonauts 2021 contains fMRIs from 10 human subjects that watched over 1,000 short (3 sec) video clips.

Other multimodal datasets. Finally, beyond the video datasets, datasets have also been proposed with other kinds of multimodality. These datasets are audiovisual ((Zinszer et al., 2018; Cao et al., 2021)), words associated with line drawings (Mitchell et al., 2008; Sudre et al., 2012), pictures along with sentences and word clouds (Pereira et al., 2018). These datasets have been collected using a variety of methods like fMRIs (Mitchell et al., 2008; Pereira et al., 2018), MEG (Sudre et al., 2012) and fNIRS (Zinszer et al., 2018; Cao et al., 2021). Specifically, in Sudre et al. (2012), subjects were asked to perform a question answering (QA) task, while their brain activity was recorded using MEG. Subjects were first presented with a question (e.g., "Is it manmade?"), followed by 60 concrete nouns, along with their line drawings, in a random order. For all other datasets, subjects performed passive viewing and/or listening.

4 Brain Regions

In this section, we discuss the mapping of brain recordings to stimulus-specific brain regions that have been discussed in the literature. Specifically, we discuss the regions of language network, auditory cortex and visual cortex, along with their sub regions. To use brain recordings from preprocessed naturalistic neuroscience datasets, follow these steps: (i) Use brain activation of voxels directly if either of the next two steps is applicable. (ii) Apply a brain mask to the brain volume to obtain the activation of voxels. or (iii) Project the brain volume onto the surface space (such as "fsaverage5," "fsaverage6," or "fsaverage"). To visualize the brain maps, popular libraries such as Nilearn [2] or Pycortex [3] are useful for fMRI recordings, while MNE-Python [4] is suitable for both MEG and EEG datasets.

[2]https://nilearn.github.io/stable/index.html
[3]https://gallantlab.org/pycortex/
[4]https://mne.tools/stable/index.html

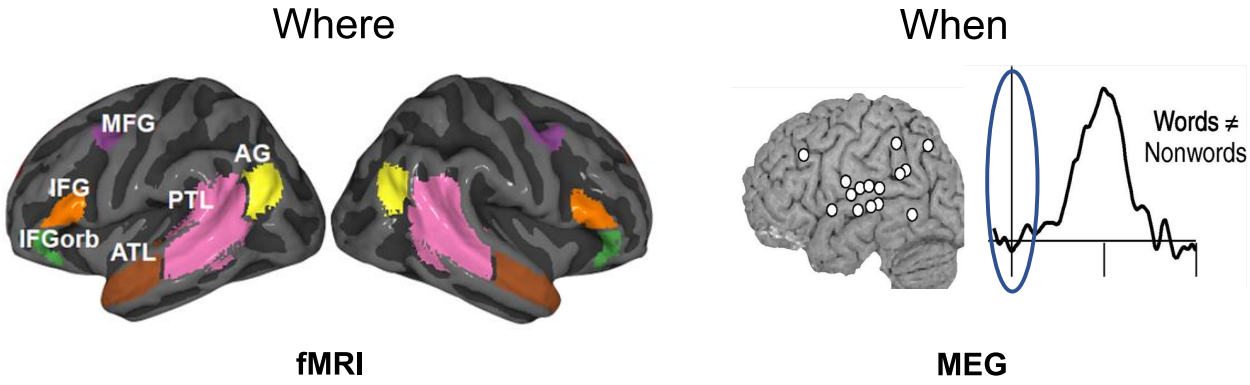


Figure 7: Non-invasive brain recordings: fMRI and MEG. This figure is adapted from (Toneva et al., 2022a).

Language network. The language network refers to brain regions that are involved in language processing. Based on the Fedorenko lab's language parcels (Fedorenko et al., 2010; Fedorenko & Thompson-Schill, 2014), eight language-relevant regions encompass broader language regions: angular gyrus (AG), anterior temporal lobe (ATL), posterior temporal lobe (PTL), inferior frontal gyrus (IFG), inferior frontal gyrus orbital (IFGOrb), middle frontal gyrus (MFG), posterior cingulate cortex (PCC) and dorsal medium prefrontal cortex (dmPFC), as shown in Figure 7 (left). These eight language networks are used in several recent studies (Toneva & Wehbe, 2019; Toneva et al., 2022a; Aw & Toneva, 2023; Oota et al., 2024b). Oota et al. (2024b; 2023d;a); Dong & Toneva (2023).

To map brain activations to these eight language regions, prior studies use the multimodal parcellation of the human cerebral cortex based on the Glasser Atlas (which consists of 180 regions of interest in each hemisphere) to report the ROI (region of interest) analysis for the brain maps (Glasser et al., 2016). Overall, the data covers eight language brain ROIs with the following subdivisions: (i) AG: PFm, PGs, PGi, TPOJ2, and TPOJ3; (ii) ATL: STSda, STSva, STGa, TE1a, TE2a, TGv, and TGd; (iii) PTL: A5, STSdp, STSvp, PSL, STV, TPOJ1; (iv) IFG: 44, 45, IFJa, IFSp; (v) MFG: 55b; (vi) IFGOrb: a47r, p47r, a9-46v, (vii) PCC: 31pv, 31pd, PCV, 7m, 23, RSC; and (viii) dmPFC: 9m, 10d, d32.

Figure 8 displays the cross-subject prediction accuracy for reading and listening for a representative sample subject. It illustrates that irrespective of text-evoked or speech-evoked brain activity, high-level information processing occurs in the language regions (indicated by white voxels).

Auditory cortex. The auditory cortex (AC) is a specific brain region responsible for processing auditory information, including the perception of sound, speech, music, and other auditory stimuli. Figure 8 displays the cross-subject prediction accuracy for reading and listening for a representative sample subject. It illustrates that during speech-evoked brain activity, the early auditory cortex (EAC) has higher prediction accuracy (indicated by Blue voxels), signifying early sensory information processing, while high-level information processing occurs in the language regions (indicated by white voxels). Overall, the auditory cortex is divided into the following subdivisions (Nastase et al., 2020): (i) EAC (early auditory cortex): A1 (Primary Auditory Cortex), the Lateral Belt (LBelt), Posterior Belt (PBelt), Medial Belt (MBelt), Rostral Intermediate (RI), and (ii) AAC (auditory association cortex): A4 and A5.

Together, these distinct areas work in concert to form a sophisticated system for perceiving and interpreting the diverse aspects of auditory stimuli.

Visual cortex. The visual cortex is a critical part of the brain responsible for visual information processing, allowing us to see and understand the world around us. Many experiments contrast brain activity elicited by specific image categories. This functional localizer approach has been used to identify many regions of interest (ROIs) in the visual pathway representing information from low-level visual (early) to high-level semantic information. The early visual cortex (EVC) is primarily responsible for processing basic visual information, including detecting of simple features like edges, colors, shapes, and motion. Higher visual

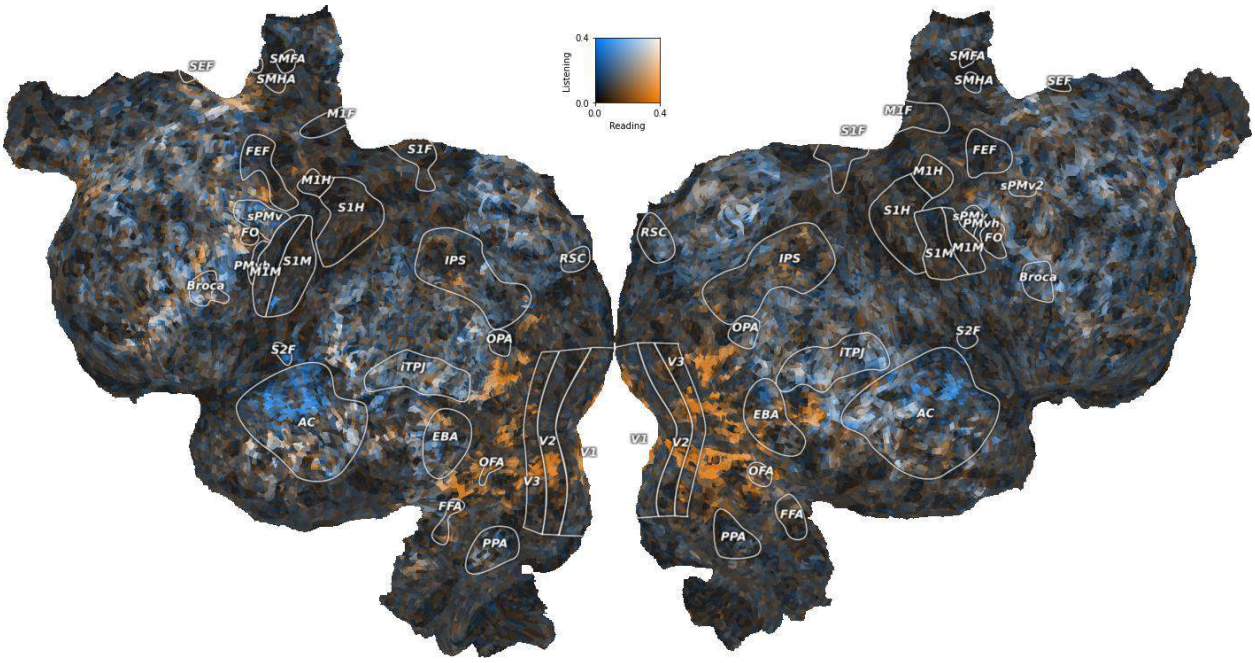


Figure 8: Contrast of estimated cross-subject prediction accuracy for reading and listening for a representative subject (subject-5). Blue and Orange voxels depict higher cross-subject prediction accuracy estimates during listening and reading, respectively. Voxels that have similar cross-subject prediction accuracy during reading and listening appear white, and are distributed across language regions. This Figure is adapted from Oota et al. (2024a).

cortex (HVC) regions are involved in more advanced visual processing tasks, such as object recognition, facial recognition, scene perception, and the integration of complex visual information.

Overall, the visual cortex is divided into the following subdivisions: (i) PVC (primary visual cortex): V1, (ii) EVC (early visual cortex): V2, V3 and V4, (ii) VWFA (visual word form area) (ii) HVC (high-level visual cortex): the extrastriate body area (EBA), occipital face area (OFA), and the fusiform face area (FFA), the occipital place area (OPA), the parahippocampal place area (PPA), and the retrosplenial cortex (RSC).

5 Evaluation Metrics

In this section, we discuss popular metrics for evaluation of brain encoding and decoding models.

5.1 Metrics for Brain Encoding Models

Two metrics are popularly used to evaluate brain encoding models: 2V2 accuracy (Toneva et al., 2020; Oota et al., 2022b) and Pearson Correlation (Jain & Huth, 2018). They are defined as follows.

Given a subject and a brain region, let N be the number of samples. Let $\{Y_i\}_{i=1}^{N}$ and $\{\hat{Y}_i\}_{i=1}^{N}$ denote the actual and predicted voxel value vectors for the i^{th} sample. Thus, $Y \in R^{N \times V}$ and $\hat{Y} \in R^{N \times V}$ where V is the number of voxels in that region.

2V2 classification accuracy. This metric evaluates how close the brain activity prediction is from ground truth, such as Euclidean distance and cosine distance. This metric evaluates the fMRI predictions using them in a classification task on held-out data in the cross-validation setting. The classification task is to try to match the predicted left-out brain responses to their corresponding ground truth, as introduced in (Mitchell et al., 2008; Wehbe et al., 2014; Toneva et al., 2020; Aw & Toneva, 2023). Having two sets of brain predictions $\hat{Y}_i$ and $\hat{Y}_j$, and corresponding ground truth Y_i and Y_j, the 2V2 classification accuracy is

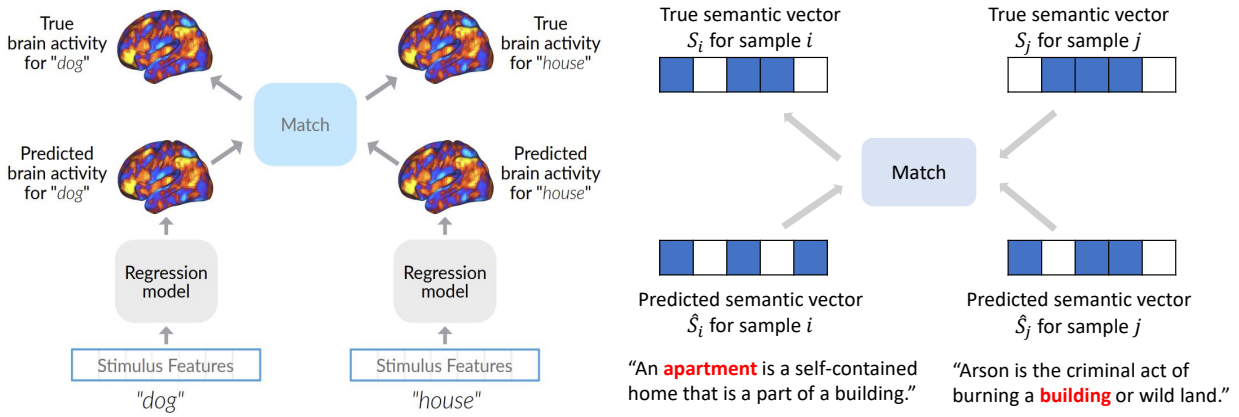


Figure 9: Evaluation Metrics for Brain Encoding and Decoding. (Left) 2V2 Accuracy (Toneva et al., 2020), (Right) Pairwise Accuracy

computed as $\frac{1}{NC_2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} I[\{\text{cosD}(Y_i,\hat{Y}_i)+\text{cosD}(Y_j,\hat{Y}_j)\} < \{\text{cosD}(Y_i,\hat{Y}_j)+\text{cosD}(Y_j,\hat{Y}_i)\}]$ where cosD is the cosine distance function. $I[c]$ is an indicator function such that $I[c] = 1$ if c is true, else it is 0. The higher the 2V2 accuracy, the better. Figure 9 (left) illustrates the computation of 2V2 Accuracy for the case where sample i and j correspond to the brain activity of concepts "dog" and "house", respectively. This metric was proposed to boost the signal-to-noise ratio in estimating the brain alignment for single-trial data (Aw & Toneva, 2023). Under this metric, chance performance is 50%.

Pearson correlation. This metric evaluates the similarity between the fMRI predictions ($\hat{Y}_i$) and the corresponding true fMRI data (Y_i) by computing the Pearson correlation for each voxel i. The Pearson correlation for voxel i is computed as $\text{PC}_i = corr[Y_i, \hat{Y}_i]$ where corr is the correlation function. The average Pearson correlation across all voxels is then computed as $\text{PCC} = \frac{1}{N}\sum_{i=1}^{n} corr[Y_i, \hat{Y}_i]$, where N denotes the number of voxels. This metric is widely used in cognitive neuroscience (Jain & Huth, 2018; Toneva & Wehbe, 2019; Caucheteux et al., 2021; Goldstein et al., 2022; Aw & Toneva, 2023; Oota et al., 2022b; 2024b).

Cross-subject prediction accuracy. To account for the intrinsic noise in biological measurements and obtain a more accurate estimate of the model's performance, Schrimpf et al. (2021) proposed an approach to estimate the cross-subject prediction accuracy. This is achieved by estimating the amount of brain response in one subject that can be predicted using only the data from a combination of other subjects using an encoding model. For instance, consider *Harry Potter* dataset with n=8 participants, the first step is to subsample—the data with n participants into all possible combinations of s participants for all $s \in [2, 8]$ (e.g. 2, 3, 4, 5, 6, 7, 8 for n=8). In the second step, for each subsample, select a random participant as the target that we attempt to predict from the remaining $s-1$ participants (e.g., predict 1 subject from 1 (other) subject, 1 from 2 subjects, ..., 1 from 8, to obtain a mean score for each voxel in that subsample. In the third step, extrapolate to infinitely many humans and thus to obtain the highest possible (most conservative) estimate, as suggested by Schrimpf et al. (2021), fit the equation $v = v_0 \times \left(1 - e^{-\frac{x}{\tau_0}}\right)$ where x is each subsample's number of participants, v is each subsample's correlation score and v_0 and τ_0 are the fitted parameters. This fitting was performed for each voxel independently with 100 bootstraps each to estimate the variance where each bootstrap draws x and v with replacement. The final ceiling value was the median of the per-voxel ceilings v_0.

Normalized brain alignment. The neural model predictivity values were normalized by their respective subject estimated cross-subject prediction accuracies, as proposed by Schrimpf et al. (2021). The final measure of a model's performance ('normalized brain alignment' or 'score') on a dataset is thus Pearson's correlation between model predictions and neural recordings divided by the estimated ceiling and averaged across voxel locations and participants.

5.2 Metrics for Brain Decoding Models

Brain decoding methods are evaluated using popular metrics like pairwise and rank accuracy (Pereira et al., 2018; Sun et al., 2019; 2020; Oota et al., 2022c). Other metrics used for brain decoding evaluation include R^2 score, mean squared error, and using Representational Similarity Matrix (Cichy et al., 2019; 2021).

Pairwise accuracy. is computed as follows. The first step is to predict all the test stimulus vector representations using a trained decoder model. Let $S = [S_0, S_1, \cdots, S_n]$, $\hat{S} = [\hat{S}_0, \hat{S}_1, \cdots, \hat{S}_n]$ denote the "true" (stimuli-derived) and predicted stimulus representations for n test instances resp. Given a pair (i, j) such that $0 \leq i, j \leq n$, score is 1 if $corr(S_i, \hat{S}_i) + corr(S_j, \hat{S}_j) > corr(S_i, \hat{S}_j) + corr(S_j, \hat{S}_i)$, else 0. Here, $corr$ denotes the Pearson correlation. Figure 9 (right) illustrates the computation of Pairwise Accuracy for the case where sample i and j correspond to the brain activations for text stimuli "apartment" and "building" respectively. Final pairwise matching accuracy per participant is the average of scores across all pairs of test instances.

Rank accuracy. is computed as follows. First, we compare each decoded vector to all the "true" stimuli-derived semantic vectors and rank them by their correlation. The classification performance reflects the rank r of the stimuli-derived vector for the correct word or picture stimuli: $1 - \frac{r-1}{\#instances-1}$. The final accuracy value for each participant is the average rank accuracy across all instances.

6 Brain Encoding

Encoding is the learning of the mapping from the stimulus domain to the neural activation. The quest in brain encoding is for "reverse engineering" the algorithms that the brain uses for sensation, perception, and higher-level cognition. The foundational approach to constructing a brain encoder, illustrated in Figure 11, adopts a general brain alignment strategy previously implemented in several notable studies (Jain & Huth, 2018; Toneva & Wehbe, 2019; Aw & Toneva, 2023; Oota et al., 2024b). This method predicts fMRI recordings at every voxel for each participant, utilizing DNN representations that mirror the participant's engagement in tasks such as reading or listening.

Building on this foundation, the recent advancements in neuroimaging technologies have enhanced our ability to closely approximate how the brain responds to different stimuli, thereby deepening our understanding of the brain's information processing mechanisms. Concurrently, advancements in deep neural network (DNN) models have led to the development of highly efficient models across different modalities, including language, vision, speech, and multimodal interactions. These models have set new benchmarks in performance for a wide range of applications. Leveraging cutting-edge neuroimaging techniques and DNN models, this section offers a comprehensive review of the task settings for brain encoding and thelatest achievements in understanding language processing, visual object recognition, auditory perception, and multimodal processing in the brain.

In the discussion on encoding task settings, we present stimulus downsampling, TR alignment, and voxelwise encoding models. In linguistic brain encoding, we explore recent breakthroughs in applied Natural Language Processing (NLP) that facilitate the reverse engineering of the language function of the brain. In the realm of vision brain encoding, pioneering results have been achieved in reverse engineering the function of the ventral visual stream for object recognition, thanks to the advancements and impressive successes of deep Convolutional Neural Networks (CNNs) and Vision Transformers. Additionally, we present the latest insights into auditory and multimodal brain encoding. This systematic approach informs the organization of this section. Overall, Figure 10 classifies the encoding literature along various stimulus domains such as vision, auditory, multimodal, and language and the corresponding tasks in each domain. Finally, Table 4 summarizes various encoding models proposed in the literature related to textual, audio, visual, and multimodal stimuli.

6.1 Encoding Task Settings

Stimulus downsampling. In the context of narrative story reading or listening, the rate of fMRI data acquisition was lower than the rate at which the text stimulus was presented to the subjects, several words fall under the same TR in a single acquisition. Hence, previous studies match the stimulus acquisition rate to

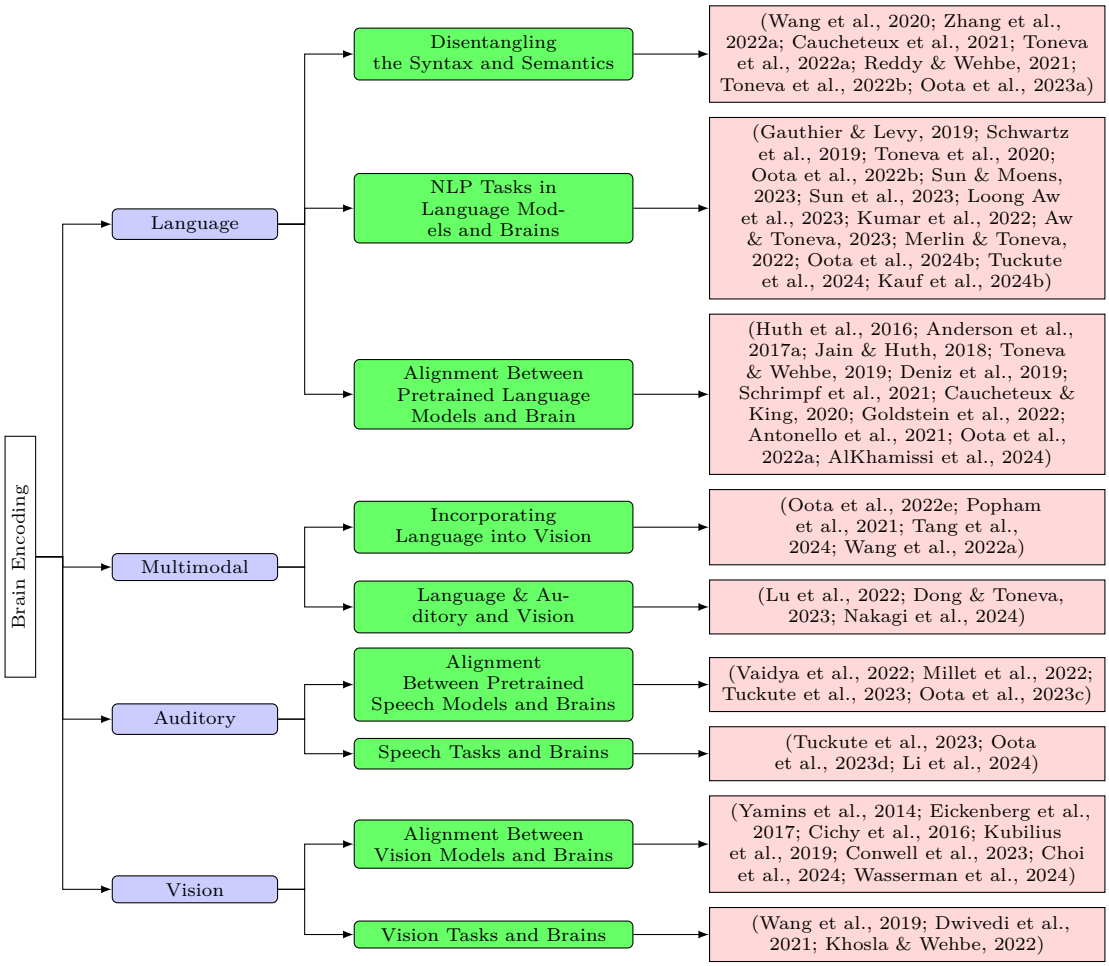


Figure 10: Categorization of Brain Encoding Studies

fMRI data recording by downsampling the stimulus features using a 3-lobed Lanczos filter (Huth et al., 2016; Jain & Huth, 2018; Toneva & Wehbe, 2019; Antonello et al., 2021; Oota et al., 2024b). After downsampling, word-embeddings corresponding to each TR are obtained.

For the naturalistic audio, (Vaidya et al., 2022; Antonello et al., 2024) windowed the stimulus waveform with a sliding window of size 16 s and stride 100ms before feeding it into the model. Further, the features are downsampled as previously described, using Lanczos interpolation, to match with sampling rate of fMRI recordings.

Similarly for the naturalistic videos, the rate of fMRI data acquisition (TR = 2 seconds) in the shortclips dataset (Huth et al., 2022) is lower than the rate at which the stimulus was presented to the subjects (15 frames per second), 30 frames of a video were viewed under the same TR for a single fMRI acquisition (Popham et al., 2021). This helps synchronization between the stimulus presentation rate and fMRI data recording, which we then leverage to train our encoding models.

fMRI Time Repetition (TR) alignment. To account for the slowness of the hemodynamic response, in general, previous studies model the HRF using a finite response filter (FIR) per voxel and for each subject separately with a delay of 8 to 12 secs (Jain & Huth, 2018; Toneva & Wehbe, 2019; Popham et al., 2021; Oota et al., 2024b; Antonello et al., 2024). Table 3 summarizes current brain encoding studies with a fixed HRF delay.

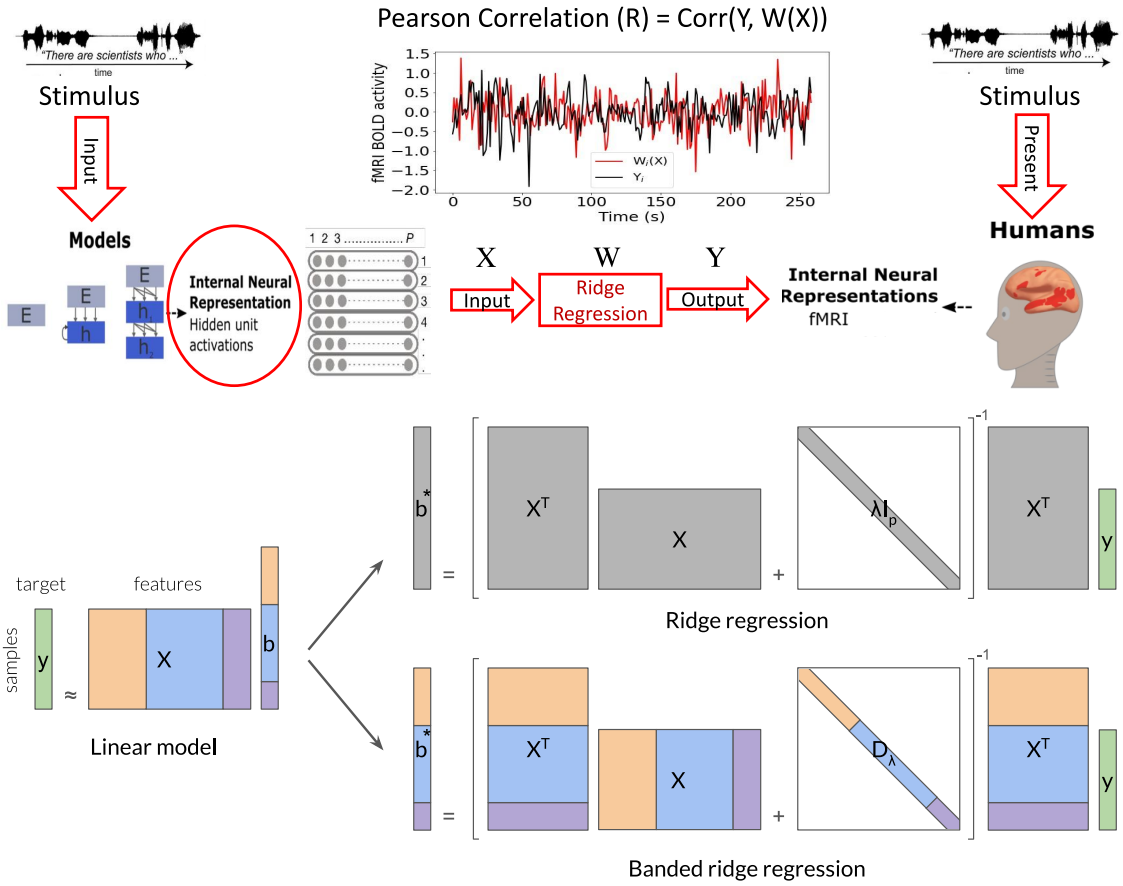


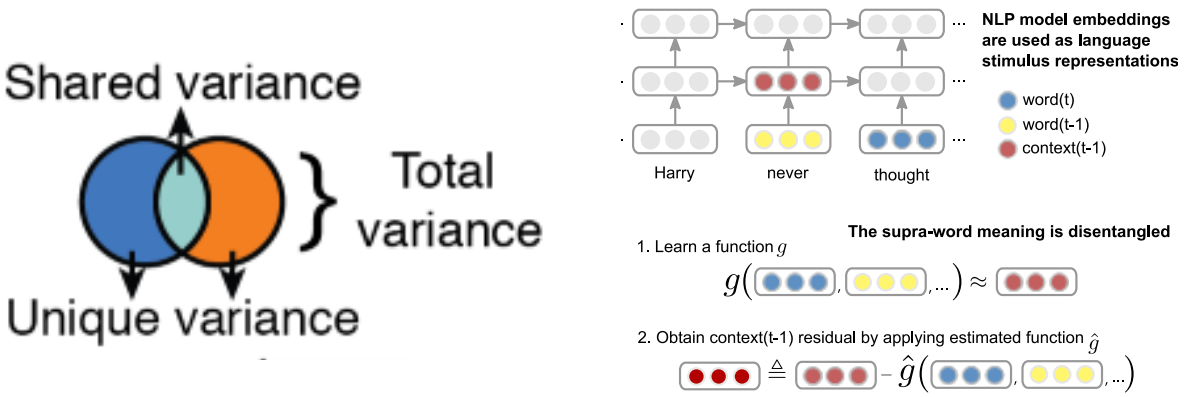
Figure 11: Scheme for Brain Encoding (top): this approach learns a function to predict the fMRI recordings at every voxel of each participant using the model representations that correspond to the same text read or listened by the participant. Ridge regression vs. Banded ridge regression (bottom) plot was *adapted from la Tour et al. (2022)*. Each color (or band) represents a different feature space.

MEG preprocessing and alignment. The minimal processing steps described in Gwilliams et al. (2023) are as follows. On raw MEG data and for each subject separately, using *MNE-Python*[5] defaults parameters, the following steps should be executed:

- bandpass filtered the MEG data between 0.5 and 30.0 Hz,
- temporally-decimated the data 10x
- segmented these continuous signals between -200 ms and 600 ms after word onset (note: this continuous signals varies for phoneme onset)
- applied a baseline correction between -200 ms and 0 ms, and
- clipped the MEG data between fifth and ninety-fifth percentile of the data across channels.

In contrast to the fMRI recordings, MEG recordings have much higher time resolution. Epoching and downsampling MEG data can result in aligned word-level or phoneme-level brain recordings (Gwilliams et al., 2023; Toneva et al., 2020; Oota et al., 2023b).

[5]https://mne.tools/stable/index.html



(a) Variance partioning

(b) Residual approach

Figure 12: Variance partitioning approach (Lescroart, 2017) (left) and Residual approach (Toneva et al., 2022a) (right).

6.2 Voxelwise Encoding Model

The main goal of the voxel-wise encoding model is to predict brain responses associated with each brain voxel given a stimulus. To estimate the brain alignment of a DNN model of stimulus representations via training standard voxel-wise encoding models Deniz et al. (2019); Toneva & Wehbe (2019). Specifically, for each voxel and participant, prior studies train fMRI encoding model using ridge regression to predict the fMRI recording associated with this voxel as a function of the stimulus representations obtained from DNN models. To simultaneously accommodate different feature spaces, which may necessitate varying levels of regularization, Nunez-Elizalde et al. (2019) proposed voxel-wise encoding model that utilize an advanced form of ridge regression. This method, known as banded ridge regression, introduces individual regularization parameters for each feature space, as illustrated in Figure 11. Before doing the ridge regression or banded ridge regression, we first z-scored each feature channel separately for training and testing. This was done to match the features to the fMRI responses, which were also z-scored for training and testing. Formally, at the time step (t), we encode the stimuli as $X_t \in \mathbb{R}^{N \times D}$ and brain region voxels $Y_t \in \mathbb{R}^{N \times V}$, where N is the number of training examples, D denotes the dimension of the concatenation of delayed TRs, and V denotes the number of voxels. To find the optimal regularization parameter for each feature space, we use a range of regularization parameters that is explored using cross-validation.

To automate the voxelwise encoding pipeline, several popular libraries have recently been introduced to perform voxelwise brain encoding: (1) Voxelwise tutorials [6] and (2) Himalaya [7]. These libraries are specifically designed for fMRI encoding models.

6.3 Interpreting brain recordings through the robustness of DNN model representations

In this section, we discuss four popular robustness methods to interpret the contribution of stimulus representations obtained from DNN models to brain alignment.

Variance partitioning. Variance partitioning quantifies the unique contribution of different stimulus features to BOLD responses. For variance partitioning (Lescroart, 2017; Deniz et al., 2019; Vaidya et al., 2022), set theory is used to calculate the common variance (as the intersection of various combinations of feature spaces) and the unique variance (as the set difference for each individual feature space), as shown in Figure 12 (a). Overall, the total variance explained by each model is computed as the unique variance explained by each model and the shared variance across models. This variance partition approach was computed separately for each voxel, then averaged across ROIs and across subjects.

[6] https://github.com/gallantlab/voxelwise_tutorials
[7] https://github.com/gallantlab/himalaya

Table 3: Summary of Brain Encoding Studies with HRF delays. Here, |S| denotes number of participants. These are studies on English text using fMRI activations.

Authors	Stimulus Representations	|S|	Dataset	Delays
Jain et al. (2020)	LSTM	6	Moth-Radio-Hour	8secs (4 TRs)
Jain & Huth (2018)	LSTM	6	Moth-Radio-Hour	8secs (4 TRs)
Caucheteux et al. (2021)	GPT-2	345	Narratives	7.5secs (5 TRs)
Reddy & Wehbe (2021)	Syntax Parsers, BERT	8	Harry-Potter	8secs (4 TRs)
Merlin & Toneva (2022)	GPT2	8	Harry-Potter	8secs (4 TRs)
Aw & Toneva (2023)	BART, LongT5, LED	8	Harry-Potter	8secs (4TRs)
Antonello et al. (2021)	100 Language Models	7	Moth-Radio-Hor	8secs (4 TRs)
Oota et al. (2024b)	BERT and Probing Tasks	18	Narratives 21st-Year	9secs (6 TRs)
Oota et al. (2023d)	BERT, GPT-2, Wav2Vec2.0	6	Moth-radio-hour	12secs (6 TRs)

Indirect approach. An indirect approach first relates model representations to the human brain, followed by an independent examination of the related model to some task performance or behavioral output. For instance, Schrimpf et al. (2021) tests the computations of a language model that may underlie human language understanding. This is accomplished by an independent examination of the relationship between the models' ability to predict an upcoming word and their brain predictivity. Similarly, Goldstein et al. (2022) provides empirical evidence that both the human brain and language model engage in continuous next-word prediction before word onset.

Residual approach. In contrast to indirect approach, the approach proposed in Toneva et al. (2022a) can directly estimate the impact of a specific feature on the alignment between the model and the brain recordings by observing the difference in alignment before and after the specific feature is computationally removed from the model representations. This method use to remove the linear contribution of a feature to a model's representation is one way to implement such a direct approach, as shown in Figure 12 (b). This is why residual approach also refer to as direct. Another method was investigated by previous work Oota et al. (2024b); Dong & Toneva (2023) and was shown to yield very similar results.

Other direct approaches have also been proposed in the literature. Most notably, work by Ramakrishnan & Deniz (2021) studies the impact of removing information related to word embeddings directly from brain responses on a downstream task. Conceptually, the results obtained from this approach and ours should be similar because the feature is completely removed from either the brain alignment input, target, or both and thus cannot further impact the observed alignment.

Stacked regression. The stacked regression approach, proposed by Lin et al. (2023), follows a two-level pipeline. The first level consists of different linear regressors, each using a different stimulus feature space as input. At the second level, the parameters α_j are learned for a convex combination of first level predictors. Overall, the entire stacked model is estimated separately at each voxel. This method is useful when building different encoding models where input feature spaces are correlated (e.g., visual and semantic features of natural images) and for demonstrating the importance of each feature space in predicting a voxel's response.

6.4 Linguistic Encoding

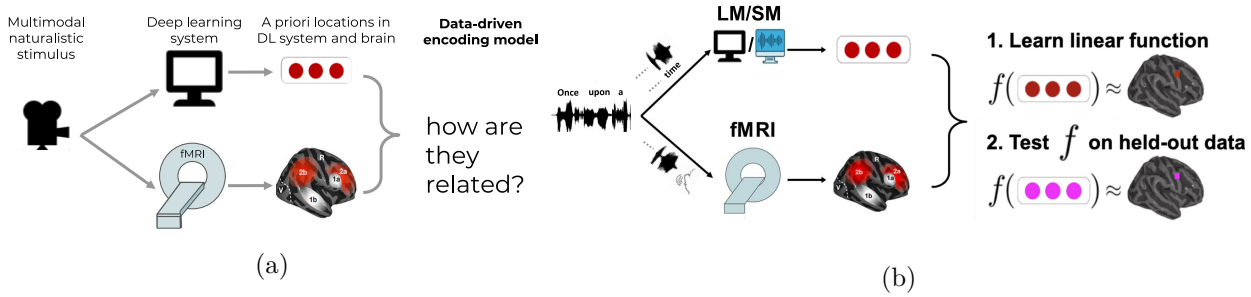
6.4.1 Alignment Between Pretrained Language Models (LMs) and Brains

Previous works have investigated the alignment between pretrained language models and brain recordings of people comprehending language. Huth et al. (2016) have been able to identify brain ROIs (Regions of Interest) that respond to words that have a similar meaning and have thus built a "semantic atlas" of how the human brain organizes language. Many studies have shown accurate results in mapping brain activity using neural distributed word embeddings for linguistic stimuli (Anderson et al., 2017a; Pereira et al., 2018; Oota et al., 2018; Nishida & Nishimoto, 2018; Sun et al., 2019). Unlike earlier models, where each word is represented as an independent vector in an embedding space, Jain & Huth (2018) built encoding models using rich contextual representations derived from an LSTM language model in a story listening task. With these contextual representations, demonstrated dissociation in brain activation – auditory cortex (AC) and Broca's area in shorter context whereas left Temporo-Parietal junction (TPJ) in longer context. Hollenstein

Table 4: Summary of Representative Brain Encoding Studies. In this table, |S| represents the number of participants in each dataset.

	Authors	Dataset Type	Lang.	Stimulus Representations	|S|	Dataset
Text	(Jain & Huth, 2018)	fMRI	English	LSTM	6	Subset Moth Radio Hour
	(Toneva & Wehbe, 2019)	fMRI, MEG	English	ELMo, BERT, Transformer-XL	9	Story understanding
	(Toneva et al., 2020)	MEG	English	BERT	9	Question-Answering
	(Schrimpf et al., 2021)	fMRI, ECoG	English	43 language models (e.g. GloVe, ELMo, BERT, GPT-2, XLNET)	20	Neural architecture of language
	(Gauthier & Levy, 2019)	fMRI	English	BERT, finetuned NLP tasks (Sentiment, Natural language inference), Scrambling language model	7	Imagine a situation with the noun
	(Deniz et al., 2019)	fMRI	English	GloVe	9	Subset Moth Radio Hour
	(Jain et al., 2020)	fMRI	English	LSTM	6	Subset Moth Radio Hour
	(Caucheteux et al., 2021)	fMRI	English	GPT-2, Basic syntax features	345	Narratives
	(Antonello et al., 2021)	fMRI	English	GloVe, BERT, GPT-2, Machine Translation, POS tasks	6	Moth Radio Hour
	(Reddy & Wehbe, 2021)	fMRI	English	Constituency, Basic syntax features and BERT	8	Harry Potter
	(Goldstein et al., 2022)	fMRI	English	GloVe, GPT-2 next word, pre-onset, post-onset word surprise	8	ECoG
	(Oota et al., 2022b)	fMRI	English	BERT and GLUE tasks	82	Pereira & Narratives
	(Oota et al., 2022a)	fMRI	English	ESN, LSTM, ELMo, Longformer	82	Narratives
	(Merlin & Toneva, 2022)	fMRI	English	BERT, Next word prediction, multi-word semantics, scrambling model	8	Harry Potter
	(Toneva et al., 2022a)	fMRI, MEG	English	ELMo, BERT, Context Residuals	8	Harry Potter
	(Aw & Toneva, 2023)	fMRI	English	BART, Longformer, Long-T5, BigBird, and corresponding Booksum models as well	8	Passive reading
	(Zhang et al., 2022b)	fMRI	English, Chinese	Node Count	19, 12	Zhang
	(Oota et al., 2023a)	fMRI	English	Constituency, Dependency trees, Basic syntax features and BERT	82	Narratives
	(Oota et al., 2023b)	MEG	English	Basic syntax features, GloVe and BERT	8	MEG-MASC
	(Tuckute et al., 2024)	fMRI	English	BERT-Large, GPT-2 XL	12	Reading Sentences
	(Kauf et al., 2024b)	fMRI	English	BERT-Large, GPT-2 XL	12	Pereira
	(Singh et al., 2023)	fMRI	English	BERT-Large, GPT-2 XL, Text Perturbations	5	Pereira
Visual	(Wang et al., 2019)	fMRI	-	21 downstream vision tasks	4	BOLD 5000
	(Kubilius et al., 2019)	fMRI	-	CNN models AlexNet, ResNet, DenseNet	7	Algonauts
	(Dwivedi et al., 2021)	fMRI	-	21 downstream vision tasks	4	BOLD 5000
	(Khosla & Wehbe, 2022)	fMRI	-	CNN models AlexNet	4	BOLD 5000
	(Conwell et al., 2023)	fMRI	-	CNN models AlexNet	4	BOLD 5000
Audio	(Millet et al., 2022)	fMRI	English	Wav2Vec2.0	345	Narratives
	(Vaidya et al., 2022)	fMRI	English	APC, AST, Wav2Vec2.0, and HuBERT	7	Moth Radio Hour
	(Tuckute et al., 2023)	fMRI	English	19 Speech Models (e.g. DeepSpeech, Wav2Vec2.0, VQ-VAE)	19	Passive listening
	(Oota et al., 2023c)	fMRI	English	5 basic and 25 deep learning based speech models (Tera, CPC, APC, Wav2Vec2.0, HuBERT, DistilHuBERT, Data2Vec	6	Moth Radio Hour
	(Oota et al., 2023d)	fMRI	English	Wav2Vec2.0 and SUPERB tasks	82	Narratives
Multi Modal	(Dong & Toneva, 2023)	fMRI	English	Merlo Reseve	5	Neuromod
	(Popham et al., 2021)	fMRI	English	985D Semantic Vector	5	Moth Radio Hour & Short Movie Clips
	(Oota et al., 2022e)	fMRI	English	CLIP, VisualBERT, LXMERT, CNNs and BERT	5, 82	Periera & Narratives
	(Lu et al., 2022)	fMRI	English	BriVL	5	Pereira & Short Movie Clips
	(Tang et al., 2024)	fMRI	English	BridgeTower	5	Moth Radio Hour & Short Movie Clips
	(Nakagi et al., 2024)	fMRI	English	BERT, GPT-2, LLaMa	5	Moth Radio Hour & Short Movie Clips

et al. (2019) presents the first multimodal framework for evaluating six types of word embeddings (Word2Vec, WordNet2Vec (Bartusiak et al., 2019), GloVe, fastText, ELMo, and BERT) on 15 datasets, including eye-tracking, EEG and fMRI signals recorded during language processing. With the recent advances in contextual representations in NLP, few studies incorporated them in relating sentence embeddings with brain activity patterns (Sun et al., 2020; Gauthier & Levy, 2019; Jat et al., 2020).



(a) (b)

Figure 13: (a) Alignment of representations between deep learning systems and human brains (Toneva & Wehbe, 2019). (b) For instance, a narrative story provided to both the Language model as well as human participants. For the Language model, we extract its representations for every word in the text. For the human participants, we record their brain activity using fMRI. Next, we train a linear function that uses the extracted Language model representations to predict human brain activity. Finally, we test this function on unseen data, and evaluate its accuracy as the amount of "brain alignment" (Toneva & Wehbe, 2019). These two images are sourced from Cogsci-22 tutorial slides Oota et al. (2022d).

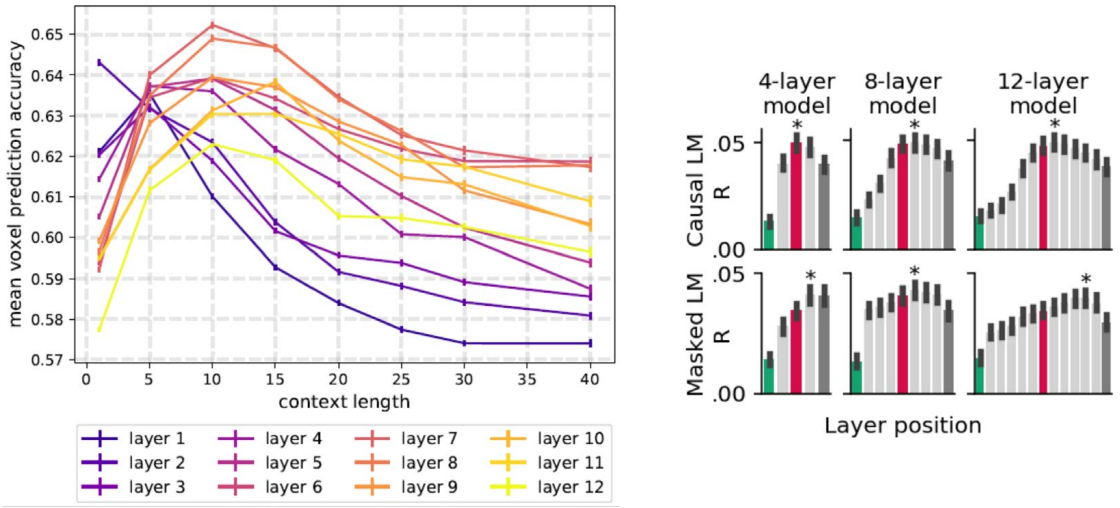


Figure 14: The strongest alignment with high-level language brain regions has consistently been observed in the middle layers. Left: Performance of BERT encoding model for all hidden layers as the amount of context provided to the network is increased (Toneva & Wehbe, 2019). Right: fMRI encoding score (averaged across time and channels) of 6 representative transformers varying in tasks (CLM vs MLM) and depth (4-12 layers) (Caucheteux & King, 2020). *The left Figure is adapted from Toneva & Wehbe (2019) and the right Figure is adapted from Caucheteux & King (2020).*

More recently, researchers have begun to study the alignment of language regions of the brain with the layers of language models (broadly following the method described in Figure 13) and found that the best alignment was achieved in the middle layers of these models (Jain & Huth, 2018; Toneva & Wehbe, 2019; Caucheteux & King, 2020), as shown in Figure 14. Toneva & Wehbe (2019) study how representations of various Transformer models differ across layer depth, context length, and attention type. The results demonstrated that across several larger NLP models, middle layers of language models are well aligned with brain language regions. Schrimpf et al. (2021) examined the relationship between 43 diverse state-of-the-art language models. They also studied the behavioral signatures of human language processing in the form of self-paced reading times and a range of linguistic functions assessed via standard engineering tasks from NLP. They found that Transformer-based models perform better than RNNs or word-level embedding models. Larger-capacity models perform better than smaller models. Models initialized with random weights

(prior to training) perform surprisingly similarly in neural predictivity compared to final trained models, suggesting that network architecture contributes as much or more than experience dependent learning to a model's match to the brain. Antonello et al. (2021) proposed a "language representation embedding space" and demonstrated the effectiveness of the features from this embedding in predicting fMRI responses to linguistic stimuli. Very recent work by (Antonello et al., 2024) tested whether larger open-source models, such as those from the text-based model (OPT and LLaMA) families, are better at predicting brain responses recorded using fMRI. The results demonstrate that encoding performance improvements scale well with both model size and dataset size, and large datasets will no doubt be necessary in producing useful encoding models.

6.4.2 Disentangling the Syntax and Semantics

The representations of transformer models like BERT and GPT-2 have been shown to linearly map onto brain activity during language comprehension. Several studies have attempted to disentangle the contributions of different types of information from word representations to the alignment between brain recordings and language models (Lopopolo et al., 2017; Wang et al., 2020; Caucheteux et al., 2021; Reddy & Wehbe, 2021; Zhang et al., 2022a; Toneva et al., 2022a; Oota et al., 2023a). Wang et al. (2020) proposed a two-channel variational autoencoder model to dissociate sentences into semantic and syntactic representations and separately associate them with brain imaging data to find feature-correlated brain regions. Similarly, Zhang et al. (2022a) separated different syntactic features from pretrained BERT representations, to explore the potential for distinct syntactic and semantic processing language regions in the brain. Compared to lexical word representations, word syntactic features (parts-of-speech, named entities) and word-relation features (semantic roles, dependencies) are distributed across brain networks instead of a local brain region. The previous two studies could not conclude whether all or any of these representations effectively drive the linear mapping between language models (LMs) and the brain. Toneva et al. (2022a) presented an approach to disentangle supra-word meaning from lexical meaning in language models and showed that supra-word meaning is predictive of fMRI recordings in two language regions (anterior and posterior temporal lobes). Similar to the approach presented in Toneva et al. (2022a), Oota et al. (2023b) disentangle the past and future context meaning from word meaning in language models and showed that past context is crucial in obtaining significant results while predicting MEG brain recordings. Caucheteux et al. (2021) proposed a taxonomy to factorize the high-dimensional activations of language models into four combinatorial classes: lexical, compositional, syntactic, and semantic representations. They found that (1) Compositional representations recruit a more widespread cortical network than lexical ones and encompass the bilateral temporal, parietal, and prefrontal cortices. (2) Contrary to previous claims, syntax and semantics are not associated with separated modules, but, instead, appear to share a common and distributed neural substrate.

While previous works studied syntactic processing as captured through complexity measures (syntactic surprisal, node count, word length, and word frequency) (Zhang et al., 2020; 2022a), very few have studied the syntactic representations themselves (Caucheteux et al., 2021; Reddy & Wehbe, 2021; Oota et al., 2023a). Studying syntactic representations using fMRI is difficult because (1) representing syntactic structure in an embedding space is a non-trivial computational problem, and (2) the fMRI signal is noisy. To overcome these limitations, Reddy & Wehbe (2021) proposed syntactic structure embeddings that encode the syntactic information inherent in the natural text that subjects read in the scanner. The results reveal that syntactic structure-based features explain additional variance in the brain activity of various parts of the language system, even after controlling for complexity metrics that capture the processing load. Toneva et al. (2022b) further examined whether the representations obtained from a language model align with different language processing regions in a similar or different way. While Reddy & Wehbe (2021) focused on constituency parsing mainly including incremental top-down parsing, Oota et al. (2023a) leverage dependency information more systematically by learning the dependency representations using graph convolutional networks, using the four step recipe as illustrated in Figure 15. The results reveal that constituency tree structure is better encoded in language regions such as bilateral temporal cortex (ATL and PTL) and MFG, while dependency structure is better encoded in AG and PCC language regions.

While previous studies focused on narrative English language stories and have shown that several brain regions are involved in building the hierarchical syntactic structure, a recent study in (Zhang et al., 2022b)

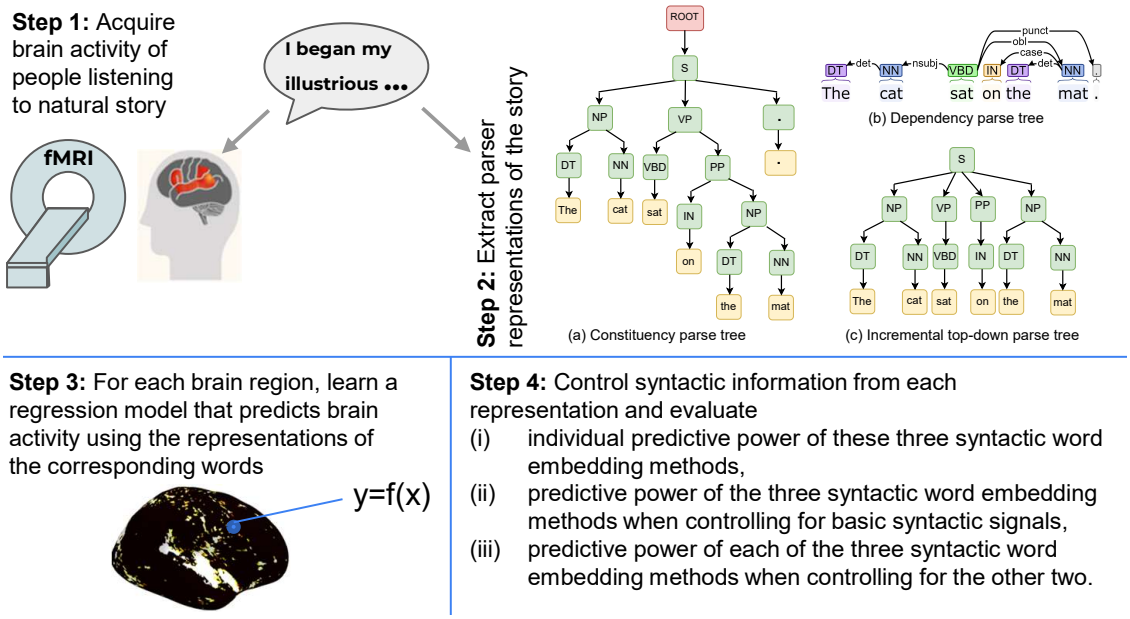


Figure 15: Four steps proposed in (Oota et al., 2023a): (1) fMRI acquisition, (2) Syntactic parsing, (3) Regression model training, and (4) Predictive power analysis of the three embeddings methods. This Figure is adapted from Oota et al. (2023a).

analyzes the neural basis of such structures between two diverse languages: Chinese and English. The results demonstrate that the brain may use different parsing strategies for different language structures to reduce the cognitive load.

6.4.3 NLP Tasks and Linguistic Properties in LMs and Brains

Understanding the reasons behind the observed similarities between language comprehension in language models and brains can lead to more insights into both systems. Further, it is unclear what type of information in the finetuned language models leads to high encoding accuracy. It is unclear whether and how the two systems align in their information processing pipeline. Recent work (Schwartz et al., 2019; Schrimpf et al., 2021; Kumar et al., 2022; Goldstein et al., 2022; Aw & Toneva, 2023; Merlin & Toneva, 2022; Oota et al., 2022b; 2024b; Sun & Moens, 2023; Sun et al., 2023; Loong Aw et al., 2023) addressed this question either by tuning the pretrained language model on downstream NLP tasks or inducing the brain relevant information into the language model.

Several researchers have suggested that one contributor to the alignment is the LM's ability to predict the next word, with a positive relationship between next-word prediction ability and brain alignment across LMs (Schrimpf et al., 2021; Goldstein et al., 2022). However, more recent work shows no simple relationship exists, and language modeling loss is not a perfect predictor of brain alignment (Pasquiou et al., 2022; Antonello et al., 2021). Schwartz et al. (2019) finetuned pretrained BERT model to predict brain activity and found that finetuned BERT has modified language representations to better encode the information that is relevant for the prediction of brain activity. Rather than finetuning BERT model on brain data, Oota et al. (2022b) finetuned BERT model on 10 GLUE (General Language Understanding Evaluation) (Wang et al., 2018) tasks to check whether task supervision leads to better encoding models to account for the brain's language representation. Oota et al. (2022b) found that using a finetuned BERT on downstream NLP tasks led to improved brain predictions. The results reveal that reading fMRI was best explained by Coreference Resolution, NER (Named Entity Recognition), shallow syntax parsing; and listening fMRI was best explained by paraphrasing, summarization, NLI. Since full finetuning generally updates the entire parameter space of the model which has been proven to distort the pretrained features (Kumar et al., 2022), Sun & Moens (2023) explore prompt-tuning that generates representations which better account for the brain's

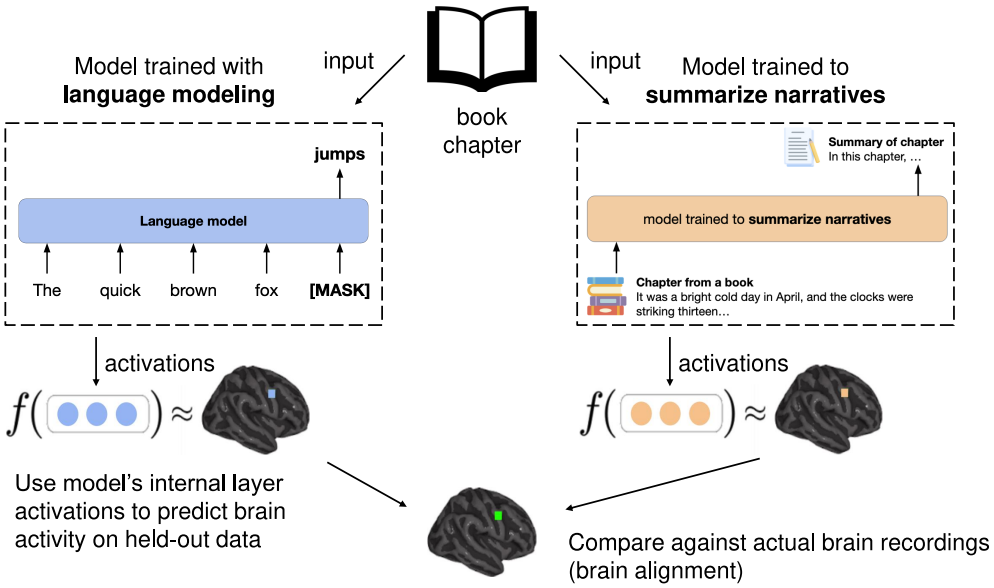


Figure 16: Comparison of brain recordings with language models trained on web corpora (Left) and language models trained on book stories (Right) (Aw & Toneva, 2023). *This Figure is redrawn from Aw & Toneva (2023).*

language representations than finetuning. They find that prompt-tuning on tasks dealing with fine-grained concept meaning including Word Sense Disambiguation and Co-reference Resolution yields representations that are better at neural decoding than tuning on other tasks with both finetuning and prompt-tuning. Further, Sun et al. (2023) extended similar prompt-tuning to bridge the gap between human brain and supervised DNN representations of the Chinese language. With the recent success of instruction-tuned large language models, Loong Aw et al. (2023) investigated the effect of instruction-tuning on large language models and alignment with the human brain's language representations. The results demonstrate that instruction-tuning of large language models (LLMs) improves both world knowledge representations and brain alignment, suggesting that mechanisms that encode world knowledge in LLMs also improve representational alignment to the human brain.

To investigate whether large language models with longer context are learning a deeper understanding of the text, Aw & Toneva (2023) used four pretrained large language models (BART, Longformer Encoder Decoder, BigBird, and LongT5) and also trained them to improve their narrative understanding, using the method detailed in Figure 16. They find that the improvements in brain alignment are larger for character names than for other discourse features, which indicates that these models are learning important narrative elements. However, it is not understood whether language models with the prediction of the next word are necessary for the observed brain alignment or simply sufficient, and whether there are other shared mechanisms or information that is similarly important. Merlin & Toneva (2022) proposed two perturbations to pretrained language models that, when used together, can control for the effects of next word prediction and word-level semantics on the alignment with brain recordings. Specifically, they found that improvements in alignment with brain recordings in two language processing regions–Inferior Frontal Gyrus (IFG) and Angular Gyrus (AG)–are due to next word prediction and word-level semantics. However, what linguistic information actually underlies the observed alignment between brains and language models was not clear. Recently, Oota et al. (2024b) tested the effect of a range of linguistic properties (surface, syntactic and semantic) and found that the elimination of each linguistic property results in a significant decrease in brain alignment across all layers of BERT. Further, syntactic properties are more responsible and have the largest effect on the trend of brain alignment across model layers. To further understand what aspects of linguistic stimuli contribute to ANN-to-brain similarity, Kauf et al. (2024b) systematically manipulated the stimuli (i.e., perturbed sentences' word order, removed different subsets of words, or replaced sentences

with other sentences of varying semantic similarity) and found that lexical semantic content rather than the sentence's syntactic form is primarily responsible for the DNN-to-brain similarity. Similar to studies on pretrained models and brain similarity, AlKhamissi et al. (2024) investigated the reasons for the similarity of untrained language models and brain alignment by performing mechanistic interpretability of the models. By isolating components of the Transformer architecture (GPT-2 XL), they found that tokenization strategy and multihead attention are the two major components driving this better brain alignment.

Previous studies (Oota et al., 2024b; Kauf et al., 2024b) on brain alignment with language models have shown mixed results, with some finding that syntactic tasks are more responsible and others emphasizing lexical semantic content. To explore this further, Kauf et al. (2024a) investigated the extent to which language comprehension relies on syntactic versus semantic cues by manipulating the grammaticality and meaningfulness of linguistic inputs. Their findings support a strong reliance on syntactic processing rather than shallow, semantics-based processing in the language network.

6.4.4 Key Takeaways

- **Alignment with Language Models:**
  1. Language models initialized with random weights (untrained models), the representations induced by architectural priors can exhibit reasonable alignment to brain data.
  2. Across several language models (like ELMo and Transformers), the middle layers of language models align well with brain language regions.
  3. Encoding performance improvements scale well with both model size and dataset size, indicating that large datasets will be essential for producing effective encoding models.
- **Semantic and Syntactic Processing:**
  1. Word syntactic and relation features are distributed across brain networks, unlike lexical word representations, which are localized to specific brain regions.
  2. Contrary to previous claims in , syntax and semantics are not associated with separate modules but instead share common brain language regions and are distributed across the language network.
- **Contextual Representations:**
  1. Brain regions like the auditory cortex and Broca's area are involved in processing shorter contexts, while regions like the left temporo-parietal junction handle longer contexts.
  2. Contextual representations from language models improve the prediction of brain activity compared to traditional word embeddings.
  3. Long past contexts enable better encoding than future or short-scale present contexts.
- **Reasons for DNN-to-Brain similarity**
  1. For untrained language models, mechanistic interpretability of models by isolating critical components of the Transformer architecture reveals that tokenization strategy and multihead attention are the two major components driving brain alignment.
  2. For pretrained language models, representational interpretability of models reveals that syntactic properties have the largest effect on the trend of brain alignment across model layers.
  3. Strong reliance of syntactic properties rather than semantic-based processing in the language network.

6.5 Auditory Encoding

To study auditory processing in the human brain, earlier studies focused on using hand-constructed features such as a number of phonemes, MFCC (Mel Frequency Cepstral Coefficients), spectrotemporal modulations for auditory brain encoding (de Heer et al., 2017). These basic acoustic features are part of a standard model of primary auditory cortex responses to sound encoding (Norman-Haignere & McDermott, 2018;

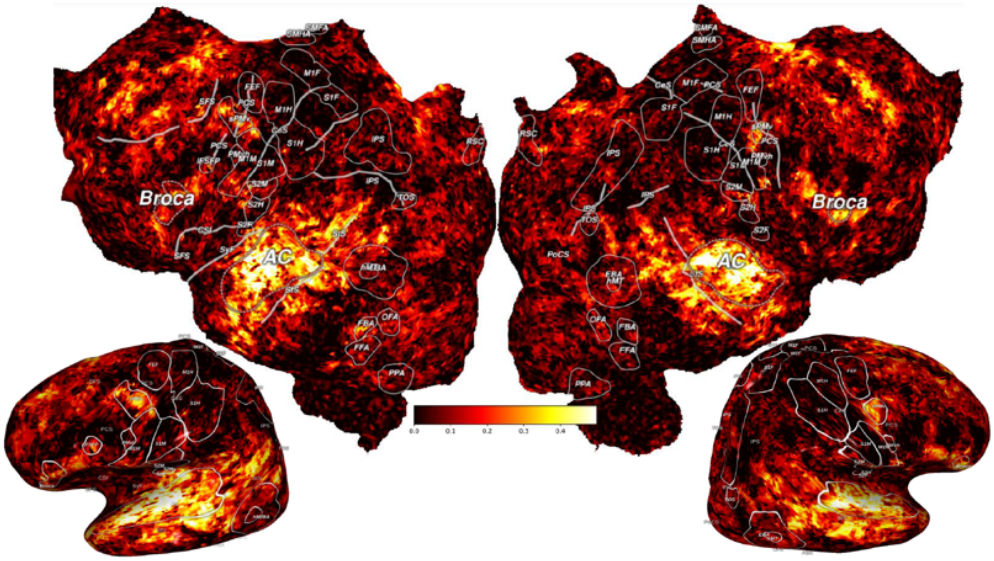


Figure 17: Brain prediction using self-supervised speech model: Data2Vec. The plot shows that speech-based models better predict early auditory cortex (Oota et al., 2023c).

Venezia et al., 2019; Mesgarani et al., 2014). In several other studies, speech stimuli have predominantly been represented as text transcriptions (Huth et al., 2016), or basic features like phoneme rate and the sum of squared FFT (Fast Fourier Transform) coefficients have been employed when constructing encoding models (Pandey et al., 2022). However, text transcription-based methods ignore the raw audio-sensory information completely. The basic speech feature engineering method misses the benefits of transfer learning from rigorously pretrained speech deep learning (DL) models. The benefits of using pretrained speech models include: (i) efficient contextual speech representations, (ii) enhanced accuracy and (iii) flexibility in fine-tuning.

6.5.1 Alignment Between Pretrained Speech Models and Brains

Recently, several researchers have used popular deep learning models such as APC (Chung et al., 2020), Wav2Vec2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and Data2Vec (Baevski et al., 2022) for encoding speech stimuli. Millet et al. (2022) used a self-supervised learning model, Wav2Vec2.0, to learn latent representations of the speech waveform similar to human brain. They find that the functional hierarchy of its transformer layers aligns with the cortical hierarchy of speech in the brain and reveals the whole-brain organisation of speech processing with unprecedented clarity. This means that the first transformer layers map onto the low-level auditory cortices (A1 and A2), the deeper layers map onto brain regions associated with higher-level processes (e.g. STS and IFG). Vaidya et al. (2022) present the first systematic study to bridge the gap between recent four self-supervised speech representation methods (APC, Wav2Vec, Wav2Vec2.0, and HuBERT) and computational models of the human auditory system. Similar to (Millet et al., 2022), they find that self-supervised speech models are the best models of auditory areas. Lower layers best modeled low-level areas, and upper-middle layers were most predictive of phonetic and semantic areas, while layer representations follow the accepted hierarchy of speech processing. Tuckute et al. (2023) analyzed 19 different speech models and found that some audio models derived in engineering contexts (model applications ranging from speech recognition and speech enhancement to audio captioning and audio source separation) produce poor predictions of auditory cortical responses, many task-optimized audio speech deep learning models outpredict a standard spectrotemporal model of the auditory cortex and exhibit hierarchical layer-region correspondence with auditory cortex. Further, Oota et al. (2023c) extended this analysis to more such deep learning based speech models (30 self-supervised speech models). They found that both language and auditory brain areas, are best aligned with intermediate layers in deep learning models. As shown in Figure 17, they also found that speech models better predict early auditory cortex than late language regions.

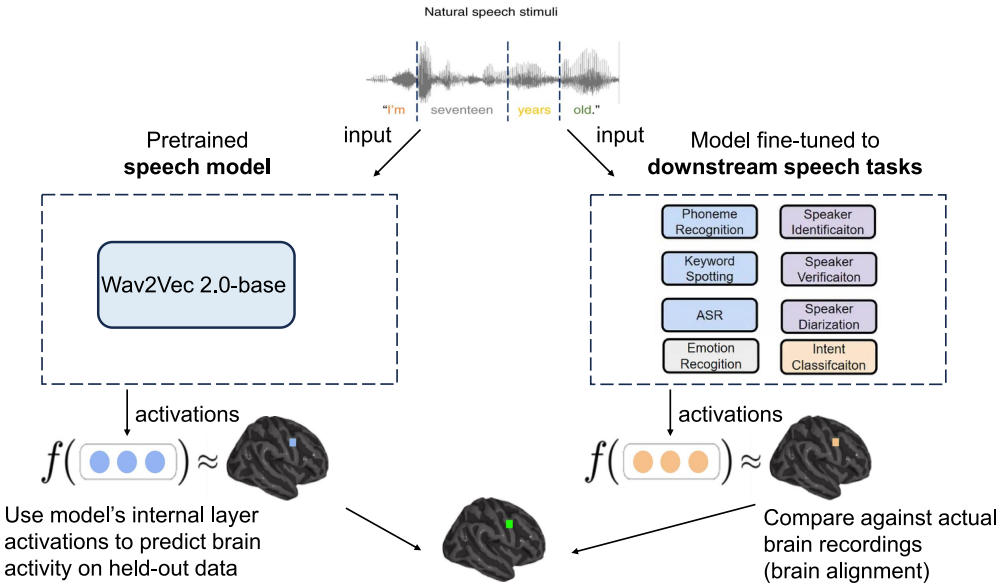


Figure 18: The pretrained Wav2Vec2.0 model and finetuned to eight different downstream speech tasks and their brain alignment (Oota et al., 2023d).

Although pretrained speech models can understand broad aspects of speech in general, the implications of finetuning speech pretrained models for various speech-processing tasks for speech encoding in the brain, remain underexplored.

6.5.2 Underlying Speech Properties in Speech Models and Brains

Understanding the reasons behind the observed similarities between speech processing in speech models and brains can lead to more insights into both systems. Recent work Oota et al. (2023d) has found that using a finetuned Wav2Vec2.0 leads to improved brain alignment. In particular, as shown in Figure 18, Oota et al. (2023d) build neural speech taskonomy models for brain encoding and aim to find speech-processing tasks that have the most explanatory capability of brain activation during naturalistic story listening experiments. They find that task-specific (Automated Speech Recognition (ASR), Entity Recognition (ER), Speaker Identification (SID) and Intent Classification (IC)) speech representations lead to a significant improvement in brain alignment compared to the pretrained Wav2Vec2.0 model for specific brain regions. Finetuning on ER, SID and IC leads to the best alignment for the early auditory cortex; finetuning on ASR provides the best encoding for the auditory associative cortex and language regions. Further, the layer-wise analysis of the effect of each speech task on the alignment with whole brain activity shows that the ASR task is better aligned in middle layers. Similar to language model fine-tuning with brain data in Schwartz et al. (2019), Li et al. (2024) fine-tuned a pretrained Wav2Vec2.0 model with brain recordings and and found that this induced process modified the language representations, improving the model's performance on downstream tasks from the SUPERB benchmark.

To understand what types of information these language models truly predict in the brain, a very recent study by Oota et al. (2024a) proposes a direct approach by removing a wide range of low-level features from model representations and examining the effect on alignment with both text and speech models. This study reveals that in context of brain reading or listening, both text-based and speech-based models show high brain alignment with late language regions, but speech models trails behind text models. In early visual and auditory regions, both models exhibit high degree of normalized brain alignment. Specifically, text models alignment with late language regions due to brain-relevant semantics, while speech models alignment due to low-level stimulus features. Conversely, text models alignment with early auditory regions mostly due to low-level textual features, while speech models alignment is only partially explained by these features. These findings conclude that speech-based language lack important brain-relevant semantics.

6.5.3 Key Takeaways

- **Alignment with Speech Models:**
  1. The functional hierarchy of Transformer layers aligns with the cortical hierarchy of speech processing in the brain.
  2. The lower layers map onto primary auditory areas, while the deeper layers are more predictive of phonetic and semantic information processing.
- **Task-specific speech models lead to improved brain alignment:**
  1. Speech tasks such as emotion recognition (ER), speaker identification (SID), and intent classification (IC) lead to the best alignment in the early auditory cortex.
  2. Fine-tuning on automatic speech recognition (ASR) provides the best alignment in the auditory association cortex and language regions.
- **Speech-based language models lack brain relevant semantics:**
  1. Speech models are useful for modeling early listening: investigate them to learn more about the auditory cortex (AC).
  2. Text models are useful for modeling late language in both listening and reading.

6.6 Visual Encoding

6.6.1 Alignment Between Vision Models and Brains

Similar to language, in vision, early models focused on independent models of visual processing (object classification) using CNNs (Yamins et al., 2014). Eickenberg et al. (2017) use CNNs as candidate models to model human brain activity during the viewing of natural images by constructing predictive models based on their different CNN layers and BOLD fMRI activations. They find that there are similarities between the computations of convolutional networks and cognitive vision at the beginning and at the end of the ventral stream object-recognition process. Cichy et al. (2016) further investigate the stages of human visual processing in both time (MEG recordings) and space (fMRI recordings). By comparing these findings with representations derived from deep neural networks (DNNs), the authors demonstrate that DNNs effectively encapsulate the sequential stages of human visual processing. This encompasses the progression from early visual areas towards the specialized pathways of the dorsal and ventral streams, highlighting the DNN's capacity to mirror complex neural processes in both time and space. Despite the effectiveness of CNNs, it is difficult to draw specific inferences about neural information processing using CNN-derived representations from a generic object-classification CNN. Hence, Wang et al. (2019) built encoding models with individual feature spaces obtained from 21 computer vision tasks. One of the main findings is that features from 3D tasks, compared to those from 2D tasks, predict a distinct part of visual cortex. Recent efforts in visual encoding models, particularly self-supervised models (instance-prototype contrastive learning), operate by taking multiple samples over an image and projecting these through a deep convolutional neural network into a low-dimensional embeddings space (Konkle & Alvarez, 2022). The results show that these self-supervised models achieve parity with the category-supervised models in accounting for the structure of brain responses. Since the human visual system uses two parallel pathways for spatial processing and object recognition, while computer vision systems (CNNs) typically use a single pathway, Choi et al. (2024) developed a dual-stream vision model to mimic human vision. This model uses two branches of CNNs to replicate the dorsal and ventral cortical pathways, aligning with the brain's pathways and suggesting that distinct responses are driven more by visual attention and object recognition goals than by retinal input selectivity.

In a recent study by Matsuyama et al. (2023) on enhancing the precision of models for visual brain encoding, the research focused on two primary questions: (1) How does changing the size of the fMRI training dataset affect prediction accuracy? (2) How does the prediction accuracy across the visual cortex change with the size of the parameters in the vision models? The findings indicate that prediction accuracy improves with increased training sample size, adhering to a scaling law. Similarly, increasing the parameter size of the vision models also leads to improved prediction accuracy, following the same scaling law.

6.6.2 Vision Tasks and Brains

How can we push deeper CNN models to capture brain processing more stringently? Continued architectural optimization on ImageNet alone no longer seems like a viable option. Instead of feed-forward deep CNN models, using shallow recurrence enabled better capture of temporal dynamics in the visual encoding models (Kubilius et al., 2019; Schrimpf et al., 2020). Kubilius et al. (2019) proposed a shallow recurrent anatomical network, CORnet, that follows neuro-anatomy more closely than standard CNNs, and achieved the state-of-the-art results on the Brain-score benchmark (Schrimpf et al., 2020). It has four computational areas, conceptualized as analogous to the ventral visual areas V1, V2, V4, and IT, and a linear category decoder that maps from the population of neurons in the model's last visual area to its behavioral choices.

6.6.3 Key Takeaways

- **Alignment with Vision Models:** The functional hierarchy of CNN layers aligns with the cortical hierarchy of visual processing in the brain.
- **Task-specific speech models lead to improved brain alignment:** Encoding models using feature spaces from 21 computer vision tasks found that features from 3D tasks predict a distinct part of the visual cortex compared to those from 2D tasks.
- **Brain-Score:** A composite of multiple neural and behavioral benchmarks is used to score any artificial neural network (ANN) based on its similarity to the brain's mechanisms for core object recognition [8].

6.7 Multimodal Brain Encoding

Recently Transformer-based multimodal models, which combine pairs of modalities such as language-vision, language-audio, and language-audio-vision, have emerged, offering rich aligned representations compared to single-modality models (i.e. text-only, audio-only or vision-only). Specifically, multimodal Transformers such as CLIP, LXMERT, and VisaulBERT take both image and text stimuli as input and output a joint visio-linguistic representations. Since the human brain perceives the environment using information from multiple modalities, examining the alignment between language and visual representations in the brain by training encoding models on fMRI responses, while extracting joint representations from multimodal models, can offer insights into the relationship between the two modalities.

Single modality stimulus. Here, participants engage in single modality stimuli, such as watching images or silent videos. Many brain encoding studies have focused on single modality stimuli, while representations are extracted from multimodal models (Oota et al., 2022e; Wang et al., 2022a; Tang et al., 2024). Oota et al. (2022e) experimented with multimodal models like CLIP, LXMERT, and VisualBERT and found VisualBERT better predict neural responses than vision-only models such as CNNs and Image Transformers. Similarly, (Wang et al., 2022a) find that multimodal models like CLIP better predict neural responses in the visual cortex than previous vision-only models like CNNs. This is attributed to the fact that high-level human visual representations encompass semantics and the relational structure of the visual world beyond object identity (Gauthier et al., 2003). Recently, Tang et al. (2024) investigated a multimodal Transformer as the encoder architecture to extract the aligned concept representations for narrative stories and movies to model fMRI responses to naturalistic stories and movies, respectively. Since language and vision rely on similar concept representations, the authors perform a cross-modal experiment in which how well the language encoding models can predict movie-fMRI responses from narrative story features (story $\rightarrow$ movie) and how well the vision encoding models can predict narrative story-fMRI responses from movie features (movie $\rightarrow$ story). Overall, the authors find that cross-modality performance was higher for features extracted from multimodal transformers than for linearly aligned features extracted from unimodal transformers.

Multimodality stimulus. Here, participants engage with multi-modal stimuli (e.g., watching movies that include audio). Recent studies have built encoding models where multi-modal stimulus representations are extracted using Transformer-based multi-modal models (Dong & Toneva, 2023; Nakagi et al., 2024). Dong

[8] https://www.brain-score.org/

& Toneva (2023) present a systematic approach to probe multimodal video Transformer model by leveraging neuro-scientific evidence of multimodal information processing in the brain. The authors find that intermediate layers of a multimodal video transformer are better at predicting multimodal brain activity than other layers, indicating that the intermediate layers encode the most brain-related properties of the video stimuli. A recent study by (Nakagi et al., 2024), which used fMRI during the viewing of 8.3 hours of video content, and discovered distinct brain regions associated with different semantic levels, highlighting the significance of modeling various levels of semantic content simultaneously. The video material was meticulously annotated in five distinct semantic categories—speech, object, story, summary, and time/place—employing advanced large language models to derive latent representations. These representations were then used to predict fMRI brain activity across the various semantic categories. The authors discovered that the lack of unique variance for Summary and TimePlace is a notable insight, suggesting that merely incorporating these types of information into encoding analyses may not adequately capture higher-level semantic representations in the brain.

6.7.1 Key Takeaways

- **Multimodal Integration:** Incorporating linguistic information with other modalities (like vision and auditory) can enhance understanding of how the brain processes complex stimuli.
- **Cross-modal vs. Jointly pretrained models:** Both cross-modal and jointly pretrained multimodal models demonstrate significantly improved brain alignment with language regions and visual regions when analyzed against unimodal video data.
- **Single modality vs. Multimodality stimulus:** Many brain encoding studies have experimented with subjects engaged with single modality stimulus, leaving the full potential of these models in true multi-modal scenarios still unclear.

7 Brain Decoding

Brain decoding aims to map neural activations back to the stimulus domain, allowing us to interpret what a person is seeing, hearing, or thinking based on their brain activity, as illustrated in Figure 19. This process is crucial for developing brain-computer interfaces and advancing our understanding of cognitive neuroscience. Unlike brain encoding, which focuses on predicting brain activity from stimuli, brain decoding involves reconstructing the original stimuli from observed neural signals (Glaser et al., 2020).

7.1 Problem Formulation

Brain decoding involves learning the mapping between brain activations and stimuli. Early approaches focused on pixel-level mappings using models such as Autoencoders (AEs) and Variational Autoencoders (VAEs), which captured detailed information but often lacked semantic richness. With the advent of large-scale generative models, the focus has shifted to conditional generation, where brain activity representations are used to condition pretrained generative models like generative adversarial networks (GANs), diffusion models, and GPTs. This shift has enhanced the fidelity and meaningfulness of decoded stimuli, enabling more sophisticated and accurate brain decoding systems.

7.2 Data Prepossessing

Similar to brain encoding, we utilized several key steps in the data preprocessing phase to ensure robust and accurate brain decoding. Initially, we performed standard preprocessing of the fMRI data, including motion correction, spatial normalization, and smoothing, to mitigate noise and artifacts inherent in the raw recordings.

We utilized paired data from previous sections, consisting of fMRI, Stimuli pairs. This paired data approach ensures that the neural activity is directly aligned with the corresponding stimuli, facilitating more accurate decoding. Following this, we extracted Regions of Interest (ROIs) based on prior neuroanatomical knowledge or functional localization tasks, ensuring that the most informative voxels were selected for subsequent

analysis. ROIs focus the analysis on specific brain areas known to be involved in processing the stimuli, reducing the dimensionality and improving the signal-to-noise ratio of the data.

7.3 Decoder Architectures

Pixel-level reconstruction. Initially, brain decoding was framed as a problem of learning an exact mapping between brain activations and stimuli, often using end-to-end models like Autoencoders (AEs) (Bank et al., 2023; Beliy et al., 2019) and Variational Autoencoders (VAEs) (Kingma & Welling, 2013; Han et al., 2019). These approaches focused on pixel-level mappings, which, while capturing detailed information, were often not semantically meaningful. Early decoding studies employed ridge regression models trained on the most informative voxels or cortex-specific voxels (Pereira et al., 2018; Sun et al., 2019; Oota et al., 2022c), with some using fully connected layers (Beliy et al., 2019) or multi-layered perceptrons (Sun et al., 2019). In some studies where decoding was modeled as multi-class classification, Gaussian Naïve Bayes (Singh et al., 2007; Just et al., 2010) and SVMs (Thirion et al., 2006) were also used. However, despite their ability to recover some detailed information (such as color, shape and location), these methods often fell short of capturing the highly complex non-linear semantic information between the stimulus and the neural responses.

Semantic reconstruction. As large-scale generative models evolved, the problem formulation shifted towards conditional generation. In this setup, a representation of brain activity is first obtained and then used as a condition for pretrained generative models, such as GANs (Du et al., 2020; Beliy et al., 2019; Fang et al., 2020), diffusion models (Chen et al., 2023; Takagi & Nishimoto, 2022; Scotti et al., 2024), and GPTs (Tang et al., 2023). This approach emphasizes learning semantic information, effectively capturing high-level information but sometimes lacking fine detail. Conditional generation models leverage vast amounts of pretrained knowledge, allowing them to generate high-quality outputs conditioned on the brain activity representations. This shift has significantly enhanced the fidelity and meaningfulness of the decoded stimuli, paving the way for more sophisticated and accurate brain decoding systems.

Trade-off between pixel-level and semantic-level reconstruction. End-to-end methods in brain decoding excel at capturing detailed information such as color, shape, and location due to their direct mapping approach from brain activations to stimuli. These models, often implemented as autoencoders or VAEs, learn a comprehensive transformation that preserves fine-grained details present in the input data. The reconstruction loss functions used during training penalize deviations from the original stimuli, encouraging the model to maintain low-level features like edges and textures.

In contrast, conditional generation frameworks involve a two-stage process where a high-level representation is first extracted from brain activity and then used to condition a pretrained generative model. While this approach leverages powerful generative models like GANs and diffusion models, which are adept at producing realistic and semantically coherent outputs, it tends to abstract away precise pixel-level details in favor of capturing broader semantic information. Consequently, end-to-end methods are particularly suited for tasks requiring detailed reconstructions, whereas conditional generation frameworks excel in generating high-level, semantically accurate representations.

Hybrid approaches. Hybrid approaches in brain decoding aim to combine the strengths of both end-to-end methods and conditional generation frameworks to achieve detailed and semantically rich reconstructions. By integrating the direct mapping capabilities of end-to-end models with the high-level semantic generation of conditional frameworks (Scotti et al., 2024; Ferrante et al., 2024; Wang et al., 2024), these approaches can capture fine-grained details while maintaining semantic coherence. Typically, a hybrid approach might first use an end-to-end model to capture detailed low-level features from brain activations and then employ a conditional generative model to refine and enhance these features, ensuring that the final output is both accurate and meaningful. This dual-stage process allows for the preservation of essential details such as color and shape while benefiting from the contextual understanding provided by advanced generative models. Hybrid approaches therefore offer a promising avenue for improving the fidelity and applicability of brain decoding technologies, bridging the gap between detailed reconstruction and high-level semantic interpretation.

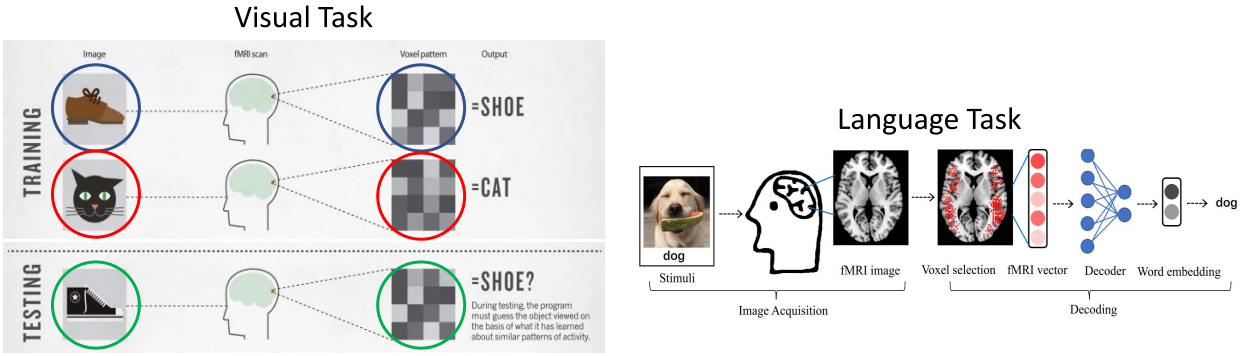


Figure 19: Scheme for Brain Decoding. Left: Image decoder (Smith, 2013), Right: Language Decoder (Wang et al., 2019). *The left Figure is adapted from Smith (2013) and the right Figure is adapted from Wang et al. (2019).*

7.4 Impact of Large Models on Brain Decoding

Representation learning. Representation learning has been a crucial step in the evolution of brain decoding. Two primary approaches have been particularly influential: masked autoencoders and contrastive learning.

Masked autoencoders (He et al., 2022) play a vital role in learning low-rank representations by reconstructing missing parts of the input data. In the context of brain decoding, these models are often used for pretraining by masking out some brain voxels and attempting to reconstruct them, thereby learning the underlying representations (Chen et al., 2023; 2024; Sun et al., 2024). These fMRI representations are then utilized as conditions for downstream conditional generation models, enhancing their ability to produce detailed and accurate reconstructions compared with linear models.

Contrastive learning (Khosla et al., 2020) has emerged as a powerful technique for representation learning by maximizing the similarity between related data points while minimizing the similarity between unrelated ones. This approach has been instrumental in aligning brain activity with corresponding stimuli in a shared embedding space, facilitating more accurate and semantically meaningful decoding. One of the most notable applications of contrastive learning in brain decoding is the CLIP model (Radford et al., 2021). CLIP aligns text and images in a shared embedding space, greatly enhancing the decoding of visual stimuli. These models decode brain activity into text descriptions that are then used to generate corresponding images, effectively bridging the gap between linguistic and visual representations. In brain decoding, researchers often align fMRI embeddings with CLIP-based embedding spaces, allowing for more precise and semantically rich reconstructions of visual stimuli from brain data Chen et al. (2024); Scotti et al. (2024).

Large language models (LLMs). LLMs, particularly models in the GPT series (Brown et al., 2020), have revolutionized language decoding. These models are capable of generating coherent and contextually appropriate text based on brain activity patterns. For instance, instead of merely decoding vector representations of stimuli, recent studies have leveraged LLMs to reconstruct entire sentences or continuous language from fMRI data Tang et al. (2023); Zhao et al. (2024). This shift from vector-based decoding to full text generation has significantly enhanced the semantic richness and contextual accuracy of the decoded output. The ability of LLMs to model complex language structures and generate text conditioned on neural data has opened new avenues for understanding how the brain processes language, providing practical applications in areas such as communication aids for individuals with speech impairments.

Diffusion models (Stable Diffusion). Diffusion models (Ho et al., 2020), particularly those like Stable Diffusion, have been pivotal in generating high-fidelity images from brain activity. These models leverage the noise-to-signal transformation process to produce detailed and semantically rich visual outputs. By conditioning these models on brain activity data, researchers have achieved remarkable success in reconstructing

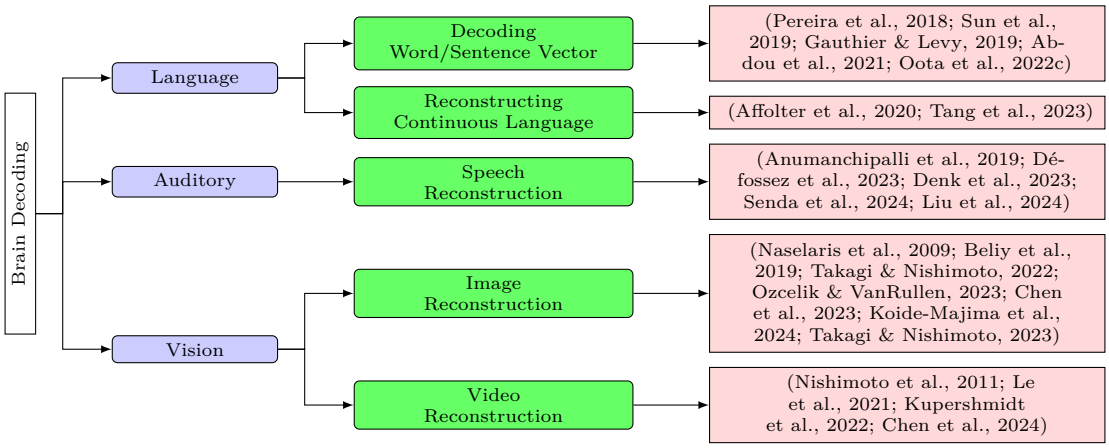


Figure 20: Categorization of Brain Decoding Studies

Table 5: Summary of Representative Brain Decoding Studies. Here, |S| represents the number of participants in each dataset.

	Authors	**Dataset Type**	**Lang.**	**Stimulus Representations**	**|S|**	**Dataset**
Text	(Pereira et al., 2018)	fMRI	English	Word2Vec, GloVe, BERT	17	Pereira
	(Wang et al., 2020)	fMRI	English	BERT, RoBERTa	6	Pereira
	(Oota et al., 2022c)	fMRI	English	GloVe, BERT, RoBERTa	17	Pereira
	(Tang et al., 2023)	fMRI	English	GPT, finetuned GPT on Reddit comments and autobiographical stories	7	Moth Radio Hour
Visual	(Beliy et al., 2019)	fMRI		End-to-End Encoder-Decoder, Decoder-Encoder, AlexNet	5	Generic Object Decoding, ViM-1
	(Takagi & Nishimoto, 2022)	fMRI		Latent Diffusion Model, CLIP	4	NSD
	(Ozcelik & VanRullen, 2023)	fMRI		VDVAE, Latent Diffusion Model	7	NSD
	(Chen et al., 2024)	fMRI		Latent Diffusion Model, CLIP	3	HCP fMRI-Video-Dataset
Audio	(Défossez et al., 2023)	MEG, EEG	English	MEL Spectrogram, Wav2Vec2.0	169	MEG-MASC
	(Gwilliams et al., 2023)	MEG	English	Phonemes	7	MEG-MASC
	(Denk et al., 2023)	fMRI	English	Music	5	Music Genre fMRI

images that closely resemble the original stimuli (Scotti et al., 2024; Takagi & Nishimoto, 2022; Chen et al., 2023; 2024; Ozcelik & VanRullen, 2023; Takagi & Nishimoto, 2023). The high resolution and fidelity of the generated images represent a significant improvement over previous methods, which often struggled to capture fine details and semantic accuracy simultaneously.

Brain decoding applications. Figure 20 summarizes the literature on decoding solutions proposed in vision, auditory, and language domains. Table 5 aggregates the brain decoding literature along different stimulus domains such as textual, visual, and audio. The most common setting is to perform decoding to a vector representation using a stimuli of a single mode (visual, text or audio).

7.5 Linguistic Decoding

Initial brain decoding experiments studied the recovery of simple concrete nouns and verbs from fMRI brain activity (Nishimoto et al., 2011) where the subject watches either a picture or a word. Sun et al. (2019) used several sentence representation models to associate brain activities with sentence stimulus, and found InferSent to perform the best. More work has focused on decoding the text passages instead of individual words (Wehbe et al., 2014). Some studies have focused on multimodal stimuli based decoding where the goal is still to decode the text representation vector. For example, Pereira et al. (2018) trained the decoder on imaging data of individual concepts, and showed that it can decode semantic vector representations from imaging data of sentences about a wide variety of both concrete and abstract topics from two separate datasets. Further, Oota et al. (2022c) propose two novel brain decoding setups: (1) multi-view decoding

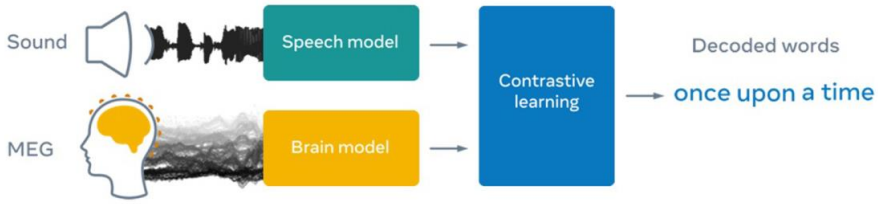


Figure 21: CLIP-MEG pipeline to align MEG activity onto pretrained speech embeddings. *The Figure is adapted from Défossez et al. (2023).*

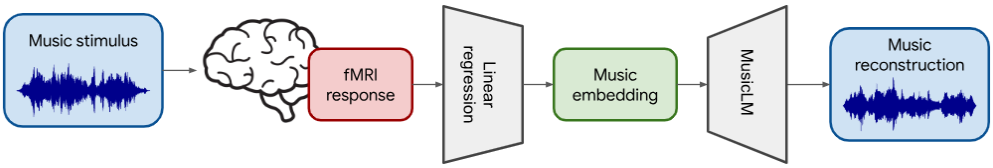


Figure 22: Brain2Music decoding pipeline adapted from (Denk et al., 2023). *The Figure is adapted from Denk et al. (2023).*

(MVD) and (2) cross-view decoding (CVD). In MVD, the goal is to build an MV decoder to take brain recordings for any view as input and predict the concept. In CVD, the goal is to train a model that takes brain recordings for one view as input and decodes a semantic vector representation of another view. Specifically, they study practically useful CVD tasks like image captioning, image tagging, keyword extraction, and sentence formation.

To understand application of Transformer models for decoding better, Gauthier & Levy (2019) finetuned a pretrained BERT on a variety of Natural Language Understanding (NLU) tasks to find tasks that lead to improvements in brain-decoding performance. They find that tasks that produce syntax-light representations (representations extracted from a language model trained on randomly shuffled words from corpus samples, thereby eliminating all first-order cues to syntactic structure) yield significant improvements in brain decoding performance.

With the recent development of large language models, rather than decoding stimuli vector representations, some studies have attempted to reconstruct words (Affolter et al., 2020), and continuous language (Tang et al., 2023) from fMRI brain activity.

7.6 Auditory Decoding

With the recent advancements of self-supervised speech models and generative AI models, recent studies have largely targeted reconstructing speech/music from brain recordings (Défossez et al., 2023; Denk et al., 2023; Senda et al., 2024). As shown in Figure 21, Défossez et al. (2023) proposed a CLIP-MEG pipeline to align MEG activity onto pretrained speech embeddings and generate speech from a stream of MEG signals. Unlike other methods which are experimented with on narrative speech, Denk et al. (2023) introduce a method for reconstructing music from fMRI brain activity, as shown in Figure 22. Specifically, they proposed a Brain2Music pipeline where the first step involves using fMRI data to predict MuLan^{music} embeddings (Huang et al., 2022), which are then passed to MusicLM (Agostinelli et al., 2023), is conditioned to generate the music reconstruction, resembling the original music stimulus.

7.7 Visual Decoding

A number of methods have been proposed for reconstructing a visual stimulus from brain recordings. Here, we initially address image reconstruction from brain recordings, followed by a discussion on video reconstruction.

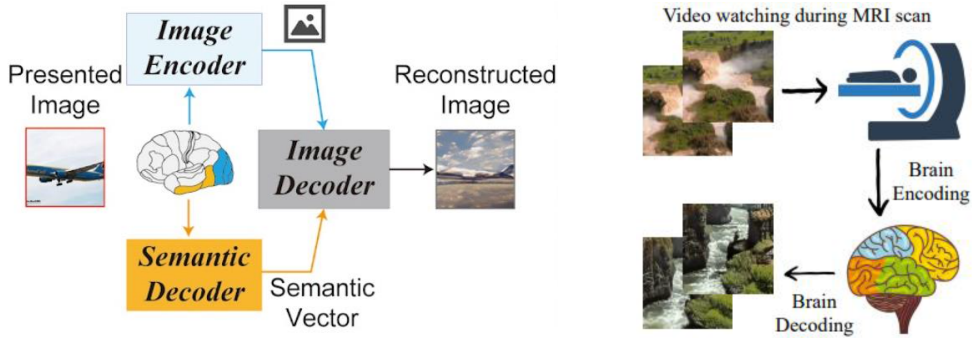


Figure 23: Image reconstruction from fMRI using Stable Diffusion. Left: (Takagi & Nishimoto, 2023), Right: (Chen et al., 2024). *The Left Figure is adapted from Takagi & Nishimoto (2023) and the Right Figure is adapted from Chen et al. (2024).*

7.7.1 Image Reconstruction

Before the success of recent generative AI models, researchers have used deep-learning models and algorithms, including generative adversarial networks (GANs) and self-supervised learning models trained on a large number of naturalistic images (Du et al., 2020; Beliy et al., 2019; Fang et al., 2020; Gaziv et al., 2022; Lin et al., 2022b). For instance, Beliy et al. (2019) designed a separable autoencoder that enables self-supervised learning in fMRI and images to increase training data. Mind Reader (Lin et al., 2022a) encoded fMRI signals into a pre-aligned vision-language latent space and used StyleGAN2 (Karras et al., 2020) for image generation. These methods generate more plausible and semantically meaningful images. Several other studies focused on reconstructing personal imagined experiences (Berezutskaya et al., 2020) or application-based decoding like using brain activity scanned during a picture-based mechanical engineering task to predict individuals' physics/engineering exam results (Cetron et al., 2019) and reflecting whether current thoughts are detailed, correspond to the past or future, are verbal or in images (Smallwood & Schooler, 2015).

With the recent success of CLIP and Diffusion models, deep generative models have been gaining attention to generate high-resolution images with high semantic fidelity (Takagi & Nishimoto, 2023; Chen et al., 2023; Scotti et al., 2024; Benchetrit et al., 2023; Song et al., 2023). (Takagi & Nishimoto, 2023) proposed a method for image reconstruction from fMRI using Stable Diffusion (Rombach et al., 2022), as shown in Figure 23(left). Their approach involves decoding brain activities to text descriptions and converting them to natural images using Stable Diffusion. Based on a similar philosophy, using a Stable Diffusion model as a generative prior and the pretrained fMRI features as conditions, (Chen et al., 2023) reconstructed high-fidelity images with high semantic correspondence to the groundtruth stimuli, as shown in Figure 23(right). (Scotti et al., 2024) proposed a MindEye that can map fMRI brain activity to any high dimensional multimodal latent space, like CLIP image space, enabling image reconstruction using generative models that accept embeddings from this latent space. Different from previous studies, BrainCLIP framework was introduced by (Liu et al., 2023) to align fMRI patterns with different modalities (especially from visual and textual modalities) through cross-modal contrastive loss. All these studies have been limited to 2D visual representations. A recent work (Gao et al., 2023) aims to extend the scope of fMRI decoding to 3D representations. Specifically, Gao et al. (2023) introduce Recon3DMind, a groundbreaking task focused on reconstructing 3D visuals from fMRI signals.

Lastly, recent image reconstruction studies have focused on other non-invasive brain recordings such as MEG and EEG rather than fMRI signals. (Benchetrit et al., 2023) proposed a CLIP-MEG pipeline to align MEG activity onto pretrained visual embeddings and generate images from a stream of MEG signals. Similarly, (Song et al., 2023) proposed a CLIP-EEG pipeline to align these two modalities (image and EEG encoders to extract features from paired image stimuli and EEG responses) by constraining their similarity.

7.7.2 Video Reconstruction

Unlike static natural images, human visual cortex can process a continuous, diverse flow of scenes, motions, and objects. To recover dynamic visual experience, the challenge lies in the nature of fMRI, which measures blood oxygenation level dependent (BOLD) signals and captures snapshots of brain activity every few seconds. Similar to image reconstruction works, Chen et al. (2024) present MinD-Video, a two-module pipeline (i.e. CLIP module followed by latent stable diffusion) designed to bridge the gap between image and video brain decoding.

7.8 Key Takeaways

- Contrastive learning models like CLIP are popular for aligning stimuli and brain data (fMRI/MEG/EEG) into a common embedding space. This alignment is useful for retrieving or reconstructing the original stimulus from brain data.
- Most decoding studies have focused on the reconstruction of stimuli such as images, videos, text, music, and speech rather than on decoding a subject's imagination. This area remains largely unexplored and would be necessary to achieve an actual mind reading label.
- Unlike brain encoding models, the interpretability of decoding models remains unexplored due to the use of more complex approaches for reconstruction. Addressing this gap is a necessary step to further examine AI models and gain deeper insights into brain functioning.

8 Conclusion, Limitations, and Future Trends

In this paper, we surveyed important naturalistic brain datasets, stimulus representations, brain encoding, and brain decoding methods across different modalities. A glimpse of how deep learning solutions throw light on putative brain computations is given. We hope that this systematic organization of recent ideas proposed in the field of cognitive computational neuroscience provides a comprehensive summary to researchers in both the AI and neuroscience communities. Insights gained from recent studies in brain encoding and decoding have significant implications for the fields of AI engineering, neuroscience, and the interpretability of models—some with immediate effects, others with long-term impact.

AI engineering. The recent brain encoding studies most immediately fit in with the neuro-AI research direction that specifically investigates the relationship between representations in the brain and representations learned by powerful neural network models. This direction has gained recent traction, especially in the domain of language, vision, and speech processing, thanks to advancements in language models (Schrimpf et al., 2021; Goldstein et al., 2022), vision models (Schrimpf et al., 2020) and speech models (Tuckute et al., 2023; Oota et al., 2023c). Furthermore, several recent works most immediately contribute to this line of research by understanding the reasons for the observed similarity in more depth (Merlin & Toneva, 2022; Oota et al., 2024b; Kauf et al., 2024b; Sarch et al., 2024; Oota et al., 2024a). Overall, these studies provide valuable insights for selecting features, enhancing transfer learning, and aiding in the creation of AI architectures that are cognitively plausible.

Computational modeling in neuroscience. Researchers have started viewing language models as useful *model organisms* for human language processing (Toneva, 2021) since they implement a language system in a way that may be very different from the human brain but may nonetheless offer insights into the linguistic tasks and computational processes that are sufficient or insufficient to solve them (McCloskey, 1991; Baroni, 2020). These brain encoding studies enable cognitive neuroscientists to have more control over using language models as model organisms of language processing. This approach can also be extended to visual and speech processing, where models in these domains serve as analogous organisms for investigation.

Model interpretability. In the long-term, we aspire for these studies on brain encoding and decoding to enhance another research direction that utilizes brain signals to interpret the information processed by neural network models (Toneva & Wehbe, 2019; Aw & Toneva, 2023; Wang et al., 2019; Sarch et al., 2024). Ultimately, our goal is to comprehend the essential and adequate underlying characteristics that result in a meaningful correlation between brain recordings and deep neural network models.

8.1 Future Trends

Some of the future areas of work in this field are as follows.

Bridging the Gap: Enhancing Deep Neural Network Models for Deeper Insights into Auditory, Language and Visual Processing While significant progress has been made in understanding text-based models, understanding the similarity in information processing between visual, speech, and multimodal models versus natural brain systems remains an open area. For instance, Oota et al. (2024a) demonstrate that speech-based language models lack brain relevant semantics in language regions. Therefore, enhancing speech-based language models to align more closely with text-based models could provide valuable insights into language and auditory processing, given that speech is the most ancient form of human language. This suggests a promising direction for future research, aiming to bridge the gap between artificial intelligence models and the complex, multifaceted processes of human cognition.

Advancing Multimodal Decoding: The Next Leap in Deep Learning Accuracy Decoding actual multimodal stimuli has become increasingly feasible due to recent advancements in deep learning models dedicated to generation tasks (Rombach et al., 2022; Singer et al., 2022). However, there is still a significant need for further research to enhance the accuracy of these models. This involves not only refining the algorithms and architectures used but also improving the quality and diversity of the datasets on which these models are trained. Advancements in computational power, algorithmic efficiency, and innovative training methodologies are critical for pushing the boundaries of what is possible in multimodal decoding, aiming to achieve more precise, reliable, and nuanced interpretations of complex stimuli.

Mapping the Mind: The Effects of Brain Damage on Cognitive Capabilities We need a deeper understanding of the degree to which damage to different regions of the human brain could lead to the degradation of selective cognitive skills. This exploration requires detailed mapping of cognitive functions to specific brain areas, taking into account the brain's complex network of connections. Studies should investigate not only the immediate effects of brain damage on cognitive skills but also the brain's capacity for reorganization and compensation over time. Ultimately, the goal is to translate these research findings into practical applications, such as more effective cognitive rehabilitation techniques and assistive technologies to improve the quality of life for individuals with brain injuries.

Towards Human-Like Understanding in ANNs: Integrating Self-Supervised Learning and Brain-Inspired Architectures How can we train artificial neural networks in novel self-supervised ways such that they compose word meanings or comprehend images and speech like a human brain? Can we model the hierarchical and modular organization of the brain in neural network architectures? This involves creating networks that reflect the brain's organization, from low-level feature detection to high-level semantic processing, allowing for the integration of information across different modalities. Moreover, how might we integrate dynamic learning strategies, such as curriculum learning, which progressively introduces more intricate tasks to the model? This method emulates how humans naturally progress from understanding straightforward to more complex ideas over time.

Bridging the Language Gap in Brain-NLP Research: The Need for Multilingual Exploration An important part of brain-NLP research relies on brain recordings collected from individuals who speak English as their primary language. Additionally, these studies utilize experimental stimuli that are presented in the English language. As a result, all current neuro-AI studies predominantly leverages language models and neural models that have been trained extensively on English text data and brain responses elicited by text or speech in English. However, it is essential to acknowledge the potential variability in our study outcomes when extrapolated to languages other than English. The intricate interplay between language-specific nuances and neural responses may introduce distinctions in the results. Therefore, it becomes imperative for future research endeavors to delve into this aspect further and investigate how these factors might influence the generalizability of our findings across diverse linguistic contexts.

In addition to the current advancements, there are several potential avenues for future exploration at the intersection of neuroscience and artificial intelligence. One such direction involves leveraging an enhanced understanding of neuroscience to propose modifications to existing artificial neural network architectures, to enhance their robustness and accuracy. Furthermore, an intriguing area for further investigation lies in

understanding the brain activity of multilingual, multi-scriptal individuals when processing stimuli in their second language (L2) or script. It remains unclear whether observed brain activity reflects the processing of L2 or the active suppression of their first language (L1) while focusing on L2. This ambiguity underscores the need for further research, particularly in the realm of multilingual multimodal stimuli, to elucidate the underlying mechanisms at play. We hope that this survey motivates research along the above directions.

References

Mostafa Abdou, Ana Valeria González, Mariya Toneva, Daniel Hershcovich, and Anders Søgaard. Does injecting linguistic structure into language models lead to better alignment with brain recordings? *arXiv preprint arXiv:2101.12608*, 2021.

Nicolas Affolter, Beni Egressy, Damian Pascual, and Roger Wattenhofer. Brain2word: Decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*, 2020.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Sarah Aliko, Jiawen Huang, Florin Gheorghiu, Stefanie Meliss, and Jeremy I Skipper. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data*, 7(1):347, 2020.

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. Brain-like language processing via a shallow untrained multihead attention network. *arXiv preprint arXiv:2406.15109*, 2024.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022.

Andrew J Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30, 2017a.

Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev DS Raizada. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, 27 (9):4379–4395, 2017b.

Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Rajeev DS Raizada, Feng Lin, and Edmund C Lalor. An integrated neural decoder of linguistic and experiential meaning. *Journal of Neuroscience*, 39(45):8969–8987, 2019.

Andrew James Anderson, Kelsey McDermott, Brian Rooks, Kathi L Heffner, David Dodell-Feder, and Feng V Lin. Decoding individual identity from brain activity elicited in imagining common experiences. *Nature Communications*, 11(1):1–14, 2020.

Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in Neural Information Processing Systems*, 34:8332–8344, 2021.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36, 2024.

Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*, 2017.

Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*, 2023.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems*, 29, 2016.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pp. 1298–1312. PMLR, 2022.

Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pp. 353–374, 2023.

Marco Baroni. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307, 2020.

Roman Bartusiak, Łukasz Augustyniak, Tomasz Kajdanowicz, Przemysław Kazienko, and Maciej Piasecki. Wordnet2vec: Corpora agnostic word vectorization method. *Neurocomputing*, 326:141–150, 2019.

Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32, 2019.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Yohann Benchetrit, Hubert Banville, and Jean-Remi King. Brain decoding: toward real-time reconstruction of visual perception. In *The Twelfth International Conference on Learning Representations*, 2023.

Julia Berezutskaya, Zachary V Freudenburg, Luca Ambrogioni, Umut Güçlü, Marcel AJ van Gerven, and Nick F Ramsey. Cortical network responses map onto data-driven features that capture visual semantics of movie fragments. *Scientific Reports*, 10(1):1–21, 2020.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Julie A Boyle, Basile Pinsard, A Boukhdhir, S Belleville, S Bram-batti, J Chen, J Cohen-Adad, A Cyr, A Fuente, P Rainville, et al. The courtois project on neuronal modelling: 2020 data release. In *OHBM*, 2020.

Jonathan R Brennan and John T Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS One*, 14(1):e0207741, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Lu Cao, Dandan Huang, Yue Zhang, Xiaowei Jiang, and Yanan Chen. Brain decoding using fnirs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12602–12611, 2021.

Charlotte Caucheteux and Jean-Rémi King. Language processing in brains and deep neural networks: computational convergence and its limits. *Nature Communications Biology*, 2020.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Disentangling syntax and semantics in the brain with deep networks. In *International Conference on Machine Learning*, pp. 1336–1348. PMLR, 2021.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 169–174, 2018.

Joshua S Cetron, Andrew C Connolly, Solomon G Diamond, Vicki V May, and James V Haxby. Decoding individual differences in stem learning from functional mri data. *Nature Communications*, 10(1):1–10, 2019.

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.

Xuhang Chen, Baiying Lei, Chi-Man Pun, and Shuqiang Wang. Brain diffuser: An end-to-end brain image to brain network pipeline. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 16–26. Springer, 2023.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36, 2024.

Minkyu Choi, Kuan Han, Xiaokai Wang, Yizhen Zhang, and Zhongming Liu. A dual-stream neural network explains the functional segregation of dorsal and ventral visual pathways in human brains. *Advances in Neural Information Processing Systems*, 36, 2024.

Yu-An Chung, Hao Tang, and James Glass. Vector-quantized autoregressive predictive coding. *Interspeech*, pp. 3760–3764, 2020.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755, 2016.

Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. In *2019 Conference on Cognitive Computational Neuroscience*. Cognitive Computational Neuroscience, 2019.

Radoslaw Martin Cichy, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Polina Iamshchinina, M Graumann, A Andonian, NAR Murty, K Kay, Gemma Roig, et al. The algonauts project 2021 challenge: How the human brain makes sense of a world in motion. *2021 Conference on Cognitive Computational Neuroscience*, 2021.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680, 2017.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, 2023. doi: 10.1101/2022.03.28.485868.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019.

Wendy A de Heer, Alexander G Huth, Thomas L Griffiths, Jack L Gallant, and Frédéric E Theunissen. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37(27):6539–6557, 2017.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023.

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736, 2019.

Timo I Denk, Yu Takagi, Takuya Matsuyama, Andrea Agostinelli, Tomoya Nakai, Christian Frank, and Shinji Nishimoto. Brain2music: Reconstructing music from human brain activity. *arXiv preprint arXiv:2307.11078*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7):431–450, 2023.

Dota Tianai Dong and Mariya Toneva. Interpreting multimodal video transformers using brain recordings. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

Changde Du, Changying Du, Lijie Huang, and Huiguang He. Conditional generative neural decoding with structured cnn feature prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2629–2636, 2020.

Kshitij Dwivedi, Michael F Bonner, Radoslaw Martin Cichy, and Gemma Roig. Unveiling functions of the visual cortex using task-specific deep neural networks. *PLOS Computational Biology*, 17(8):e1009267, 2021.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

Tao Fang, Yu Qi, and Gang Pan. Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems*, 33:13038–13048, 2020.

Evelina Fedorenko and Sharon L Thompson-Schill. Reworking the language network. *Trends in Cognitive Sciences*, 18(3):120–126, 2014.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194, 2010.

Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Through their eyes: multi-subject brain decoding with simple alignment techniques. *Imaging Neuroscience*, 2:1–21, 2024.

Jianxiong Gao, Yuqian Fu, Yun Wang, Xuelin Qian, Jianfeng Feng, and Yanwei Fu. Mind-3d: Reconstruct high-quality 3d objects in human brain. *arXiv preprint arXiv:2312.07485*, 2023.

Isabel Gauthier, Thomas W James, Kim M Curby, and Michael J Tarr. The influence of conceptual knowledge on visual discrimination. *Cognitive Neuropsychology*, 20(3-6):507–523, 2003.

Jon Gauthier and Roger Levy. Linking artificial and human neural representations of language. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254:119121, 2022.

Joshua I Glaser, Ari S Benjamin, Raeed H Chowdhury, Matthew G Perich, Lee E Miller, and Konrad P Kording. Machine learning for neural decoding. *eneuro*, 7(4), 2020.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. Random walks and neural network language models on knowledge bases. In *Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1434–1439, 2015.

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.

Zhengxin Gong, Ming Zhou, Yuxuan Dai, Yushan Wen, Youyi Liu, and Zonglei Zhen. A large-scale fmri dataset for the visual processing of naturalistic scenes. *Scientific Data*, 10(1):559, 2023.

Mandy Guo, Joshua Ainslie, David C Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. Longt5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 724–736, 2022.

Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. Introducing meg-masc a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific Data*, 10(1):862, 2023.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2727–2736, 2018.

Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, 2019.

Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Alessandro Lenci, Luca Cecchetti, Mirco Cosottini, and Giovanna Marotta. How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *NeuroImage*, 135:232–242, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(1): 1–13, 2018.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. Cognival: A framework for cognitive word embedding evaluation. In *CoNLL*, pp. 538–549, 2019.

Tomoyasu Horikawa. Mind captioning: Evolving descriptive text of mental content from human brain activity. *bioRxiv*, pp. 2024–04, 2024.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):1–15, 2017.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *TASLP*, 29:3451–3460, 2021.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. Mulan: A joint embedding of music audio and natural language. In *Ismir 2022 Hybrid Conference*, 2022.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

Alexander G Huth, Shinji Nishimoto, An T Vu, Dupre la Tour T, and Gallant JL. Gallant lab natural short clips 3t fmri data. *G-Node*, 2022. URL `https://doi.org/10.12751/g-node.vy1zjd`.

Anna A Ivanova, Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. Beyond linear regression: mapping models in cognitive neuroscience should align with research goals. *Neurons, Behavior, Data analysis, and Theory*, 5, 2022. doi: https://doi.org/10.51628/001c.37507.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in Neural Information Processing Systems*, 31, 2018.

Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander Huth. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. *Advances in Neural Information Processing Systems*, 33:13738–13749, 2020.

S Jat, H Tang, P Talukdar, and T Mitchel. Relating simple sentence representations in deep neural networks and the brain. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 5137–5154, 2020.

Heejung Jung, Maryam Amini, Bethany J Hunt, Eilis I Murphy, Patrick Sadil, Yaroslav O Halchenko, Bogdan Petre, Zizhuang Miao, Philip A Kragel, Xiaochun Han, et al. A multimodal fmri dataset unifying naturalistic processes with a rich array of experimental tasks. *bioRxiv*, pp. 2024–06, 2024.

Marcel Adam Just, Vladimir L Cherkassky, Sandesh Aryal, and Tom M Mitchell. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLOS One*, 5(1):e8622, 2010.

Antonia Karamolegkou, Mostafa Abdou, and Anders Søgaard. Mapping brains with language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 9748–9762, 2023.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Carina Kauf, Hee So Kim, Elizabeth J Lee, Niharika Jhingan, Jingyuan Selena She, Maya Taliaferro, Edward Gibson, and Evelina Fedorenko. Linguistic inputs must be syntactically parsable to fully engage the language network. *bioRxiv*, pp. 2024–06, 2024a.

Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. Lexical-semantic content, not syntactic structure, is the main contributor to ann-brain similarity of fmri responses in the language network. *Neurobiology of Language*, 5(1):7–42, 2024b.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

Meenakshi Khosla and Leila Wehbe. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, 2022.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. *Advances in Neural Information Processing Systems*, 28, 2015.

Naoko Koide-Majima, Shinji Nishimoto, and Kei Majima. Mental image reconstruction from human brain activity: Neural decoding of mental imagery via deep neural network-based bayesian estimation. *Neural Networks*, 170:349–363, 2024.

Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, 13(1):491, 2022.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in Neural Information Processing Systems*, 32, 2019.

Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *Nature Communications*, pp. 2022–06, 2022.

Ganit Kupershmidt, Roman Beliy, Guy Gaziv, and Michal Irani. A penny for your (visual) thoughts: Self-supervised reconstruction of natural movies from brain activity. *arXiv preprint arXiv:2206.03544*, 2022.

Tom Dupré la Tour, Michael Eickenberg, Anwar O Nunez-Elizalde, and Jack L Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022.

Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Bold moments: modeling short visual events through a video fmri dataset and metadata. *bioRxiv*, pp. 2023–03, 2023.

Lynn Le, Luca Ambrogioni, Katja Seeliger, Yağmur Güçlütürk, Marcel van Gerven, and Umut Güçlü. Brain2pix: Fully convolutional naturalistic video reconstruction from brain activity. *BioRxiv*, pp. 2021–02, 2021.

Submitter Mark Lescroart. Unique variance found by variance partitioning is superior to total variance explained as a model comparison metric. *Cognitive Computational Neuroscience*, 2017.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Hengyu Li, Kangdi Mei, Zhaoci Liu, Yang Ai, Liping Chen, Jie Zhang, and Zhenhua Ling. Refining self-supervised learnt speech representation using brain activations. *arXiv preprint arXiv:2406.08266*, 2024.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R Nathan Spreng, Jonathan R Brennan, Yiming Yang, Christophe Pallier, and John Hale. Le petit prince: A multilingual fmri corpus using ecological stimuli. *Scientific Data*, pp. 2021–10, 2021.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5265–5275, 2020.

Ruogu Lin, Thomas Naselaris, Kendrick Kay, and Leila Wehbe. Stacked regressions and structured variance partitioning for interpretable brain maps. *bioRxiv*, 2023.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35:29624–29636, 2022a.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35:29624–29636, 2022b.

Che Liu, Changde Du, Xiaoyu Chen, and Huiguang He. Reverse the auditory processing pathway: Coarse-to-fine audio reconstruction from fmri. *arXiv preprint arXiv:2405.18726*, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yulong Liu, Yongqiang Ma, Wei Zhou, Guibo Zhu, and Nanning Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding from fmri. *arXiv preprint arXiv:2302.12971*, 2023.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain. *arXiv e-prints*, pp. arXiv–2312, 2023.

Alessandro Lopopolo, Stefan L Frank, Antal Van den Bosch, and Roel M Willems. Using stochastic language models (slm) to map lexical, syntactic, and phonological information processing in the brain. *PLOS One*, 12(5):e0177794, 2017.

Alessandro Lopopolo, Stefan L Frank, Antal Van den Bosch, Annabel Nijhof, and Roel M Willems. The narrative brain dataset (nbd), an fmri dataset for the study of natural language processing in the brain. *Linguistic and Neuro-Cognitive Resources (LiNCR)*, 2018.

Haoyu Lu, Qiongyi Zhou, Nanyi Fei, Zhiwu Lu, Mingyu Ding, Jingyuan Wen, Changde Du, Xin Zhao, Hao Sun, Huiguang He, et al. Multimodal foundation models are better simulators of the human brain. *arXiv preprint arXiv:2208.08263*, 2022.

Takuya Matsuyama, Kota S Sasaki, and Shinji Nishimoto. Applicability of scaling laws to vision encoding models. *arXiv preprint arXiv:2308.00678*, 2023.

Michael McCloskey. Networks and theories: The place of connectionism in cognitive science. *Psychological science*, 2(6):387–395, 1991.

Gabriele Merlin and Mariya Toneva. Language models and brain alignment: beyond word-level semantics and prediction. *arXiv preprint arXiv:2212.00596*, 2022.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi King, et al. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35:33428–33443, 2022.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, and Robert A Mason. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880): 1191–1195, 2008.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. The brain tells a story: Unveiling distinct representations of semantic content in speech, objects, and stories in the human brain with large language models. *bioRxiv*, pp. 2024–02, 2024.

Tomoya Nakai, Naoko Koide-Majima, and Shinji Nishimoto. Music genre neuroimaging dataset. *Data in Brief*, 40:107675, 2022.

Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

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. Narratives: fmri data for evaluating models of naturalistic language comprehension. preprint. *Neuroscience, December*, pp. 2020–06, 2020.

Satoshi Nishida and Shinji Nishimoto. Decoding naturalistic experiences from human brain activity via distributed representations of words. *NeuroImage*, 180:232–242, 2018.

Satoshi Nishida, Yusuke Nakano, Antoine Blanc, Naoya Maeda, Masataka Kado, and Shinji Nishimoto. Brain-mediated transfer learning of convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5281–5288, 2020.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.

Sam V Norman-Haignere and Josh H McDermott. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLOS Biology*, 16(12):e2005127, 2018.

Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, 197:482–492, 2019.

Subba Reddy Oota, Naresh Manwani, and Raju S Bapi. fMRI Semantic Category Decoding Using Linguistic Encoding of Word Embeddings. In *ICONIP*, pp. 3–15. Springer, 2018.

Subba Reddy Oota, Vijay Rowtula, Manish Gupta, and Raju S Bapi. Stepencog: A convolutional lstm autoencoder for near-perfect fmri encoding. In *IJCNN*, pp. 1–8. IEEE, 2019.

Subba Reddy Oota, Frederic Alexandre, and Xavier Hinaut. Long-term plausibility of language models and neural dynamics during narrative listening. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022a.

Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. Neural language taskonomy: Which nlp tasks are the most predictive of fmri brain activity? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3220–3237, 2022b.

Subba Reddy Oota, Jashn Arora, Manish Gupta, and Raju S Bapi. Multi-view and cross-view brain decoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 105–115, 2022c.

Subba Reddy Oota, Jashn Arora, Manish Gupta, Raju Surampudi Bapi, and Mariya Toneva. Deep learning for brain encoding and decoding. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022d.

Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S Bapi. Visio-linguistic brain encoding. In *COLING 2022-the 29th International Conference on Computational Linguistics*, pp. 116–133, 2022e.

Subba Reddy Oota, Mounika Marreddy, Manish Gupta, and Raju Bapi. How does the brain process syntactic structure while listening? In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6624–6647, 2023a.

Subba Reddy Oota, Trouvain Nathan, Frederic Alexandre, and Xavier Hinaut. Meg encoding using word context semantics in listening stories. In *Interspeech*, 2023b.

Subba Reddy Oota, Khushbu Pahwa, Mounika Marreddy, Manish Gupta, and Raju Surampudi Bapi. Neural architecture of speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023c.

Subba Reddy Oota, Agarwal Veeral, Marreddy Mounika, Gupta Manish, and Raju Surampudi Bapi. Speech taskonomy: Which speech tasks are the most predictive of fmri brain activity? In *Interspeech*, 2023d.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. *Association for Computational Linguistics: ACL 2024*, 2024a.

SubbaReddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36, 2024b.

Yohei Oseki and M Asahara. Design of bccwj-eeg: Balanced corpus with human electroencephalography. In *LREC*, pp. 189–194, 2020.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

Pankaj Pandey, Gulshan Sharma, Krishna P Miyapuram, Ramanathan Subramanian, and Derek Lomas. Music identification using brain responses to initial snippets. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1246–1250, 2022.

Alexandre Pasquiou, Yair Lakretz, John T Hale, Bertrand Thirion, and Christophe Pallier. Neural language models are not born equal to fit brain data, but training helps. In *International Conference on Machine Learning*, pp. 17499–17516. PMLR, 2022.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Francisco Pereira, Matthew Botvinick, and Greg Detre. Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial Intelligence*, 194:240–252, 2013.

Francisco Pereira, Bin Lou, Brianna Pritchett, Nancy Kanwisher, Matthew Botvinick, and Ev Fedorenko. Decoding of generic mental representations from functional mri data using word embeddings. *bioRxiv*, pp. 057216, 2016.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):1–13, 2018.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1499–1509, 2018.

Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11):1628–1636, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

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.

Kalyan Ramakrishnan and Fatma Deniz. Non-complementarity of information in word-embedding and brain representations in distinguishing between concrete and abstract words. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 1–11, 2021.

Aniketh Janardhan Reddy and Leila Wehbe. Can fmri reveal the representation of syntactic structure in the brain? *Advances in Neural Information Processing Systems*, 34:9843–9856, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Gabriel Sarch, Michael Tarr, Katerina Fragkiadaki, and Leila Wehbe. Brain dissection: fmri-trained networks reveal spatial selectivity in the processing of natural images. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

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), 2021.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. *Advances in Neural Information Processing Systems*, 32, 2019.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36, 2024.

K Seeliger, RP Sommers, Umut Güçlü, Sander E Bosch, and MAJ Van Gerven. A large single-participant fmri dataset for probing brain responses to naturalistic stimuli in space and time. *bioRxiv*, pp. 687681, 2019.

Jyun Senda, Mai Tanaka, Keiya Iijima, Masato Sugino, Fumina Mori, Yasuhiko Jimbo, Masaki Iwasaki, and Kiyoshi Kotani. Auditory stimulus reconstruction from ecog with dnn and self-attention modules. *Biomedical Signal Processing and Control*, 89:105761, 2024.

K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2022.

Chandan Singh, Aliyah R Hsu, Richard Antonello, Shailee Jain, Alexander G Huth, Bin Yu, and Jianfeng Gao. Explaining black box text modules in natural language with language models. *XAI in Action: Past, Present, and Future Applications, NeurIPS-2023 Workshop*, 2023.

Vishwajeet Singh, Krishna P. Miyapuram, and Raju S. Bapi. Detection of cognitive states from fmri data using machine learning techniques. In Manuela M. Veloso (ed.), *IJCAI*, pp. 587–592, 2007.

Jonathan Smallwood and Jonathan W Schooler. The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology*, 66:487–518, 2015.

Kerri Smith. Reading minds. *Nature*, 502(7472):428, 2013.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. In *The Twelfth International Conference on Learning Representations*, 2023.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*, 2018.

Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62 (1):451–463, 2012.

Jingyuan Sun and Marie-Francine Moens. Fine-tuned vs. prompt-tuned supervised representations: which better account for brain language representations? In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 5197–5205, 2023.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Towards sentence-level brain decoding with distributed representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7047–7054, 2019.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Neural encoding and decoding with distributed sentence representations. *IEEE TNNLS*, 32(2):589–603, 2020.

Jingyuan Sun, Xiaohan Zhang, and Marie-Francine Moens. Tuning in to neural encoding: Linking human brain and artificial supervised representations of language. In *ECAI 2023*, pp. 2258–2265. IOS Press, 2023.

Jingyuan Sun, Mingxiao Li, Zijiao Chen, Yunhao Zhang, Shaonan Wang, and Marie-Francine Moens. Contrast, attend and diffuse to decode high-resolution images from brain activities. *Advances in Neural Information Processing Systems*, 36, 2024.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *Conference on Computer Vision and Pattern Recognition*, pp. 2022–11, 2022.

Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. *arXiv preprint arXiv:2306.11536*, 2023.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, 2019.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

Jerry Tang, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems*, 36, 2024.

Qawi K Telesford, Eduardo Gonzalez-Moreira, Ting Xu, Yiwen Tian, Stanley J Colcombe, Jessica Cloud, Brian E Russ, Arnaud Falchier, Maximilian Nentwich, Jens Madsen, et al. An open-access dataset of naturalistic viewing using simultaneous eeg-fmri. *Scientific Data*, 10(1):554, 2023.

Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4):1104–1116, 2006.

Mariya Toneva. *Bridging Language in Machines with Language in the Brain.* PhD thesis, Carnegie Mellon University, 2021.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 32, 2019.

Mariya Toneva, Otilia Stretcu, Barnabás Póczos, Leila Wehbe, and Tom M Mitchell. Modeling task effects on meaning representation in the brain via zero-shot meg prediction. *Advances in Neural Information Processing Systems*, 33:5284–5295, 2020.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2(11):745–757, 2022a.

Mariya Toneva, Jennifer Williams, Anand Bollu, Christoph Dann, and Leila Wehbe. Same cause; different effects in the brain. In *First Conference on Causal Learning and Reasoning*, pp. 787–825. MLR Press, 2022b.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biology*, 21(12):e3002366, 2023.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024.

Aditya R Vaidya, Shailee Jain, and Alexander Huth. Self-supervised models of audio effectively explain human cortical responses to speech. In *International Conference on Machine Learning*, pp. 21927–21944. PMLR, 2022.

Marcel Van Gerven, Jason Farquhar, Rebecca Schaefer, Rutger Vlek, Jeroen Geuze, Anton Nijholt, Nick Ramsey, Pim Haselager, Louis Vuurpijl, Stan Gielen, et al. The brain–computer interface cycle. *Journal of Neural Engineering*, 6(4):041001, 2009.

Jonathan H Venezia, Steven M Thurman, Virginia M Richards, and Gregory Hickok. Hierarchy of speech-driven spectrotemporal receptive fields in human auditory cortex. *NeuroImage*, 186:647–666, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.

Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *Advances in Neural Information Processing Systems*, 32, 2019.

Aria Yuan Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *BioRxiv*, pp. 2022–09, 2022a.

Jing Wang, Vladimir L Cherkassky, and M Adam Just. Predicting the brain activation pattern associated with the propositional content of a sentence: Modeling neural representations of events and states. *Human Brain Mapping*, 10:4865–4881, 2017.

Shaonan Wang, Jiajun Zhang, Haiyan Wang, Nan Lin, and Chengqing Zong. Fine-grained neural decoding with distributed word representations. *Information Sciences*, 507:256–272, 2020.

Shaonan Wang, Xiaohan Zhang, Jiajun Zhang, and Chengqing Zong. A synchronized multimodal neuroimaging dataset for studying brain language processing. *Scientific Data*, 9(1):590, 2022b.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11333–11342, 2024.

Navve Wasserman, Roman Beliy, Roy Urbach, and Michal Irani. Functional brain-to-brain transformation with no shared data. *arXiv preprint arXiv:2404.11143*, 2024.

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, 2014.

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.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

Xiaohan Zhang, Shaonan Wang, Nan Lin, Jiajun Zhang, and Chengqing Zong. Probing word syntactic representations in the brain by a feature elimination method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11721–11729, 2022a.

Xiaohan Zhang, Shaonan Wang, Nan Lin, and Chengqing Zong. Is the brain mechanism for hierarchical structure building universal across languages? an fmri study of chinese and english. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7852–7861, 2022b.

Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications*, 11(1):1–13, 2020.

Xinpei Zhao, Jingyuan Sun, Shaonan Wang, Jing Ye, Xiaohan Zhang, and Chengqing Zong. Mapguide: A simple yet effective method to reconstruct continuous language from brain activities. *arXiv preprint arXiv:2403.17516*, 2024.

Benjamin D Zinszer, Laurie Bayet, Lauren L Emberson, Rajeev DS Raizada, and Richard N Aslin. Decoding semantic representations from functional near-infrared spectroscopy signals. *Neurophotonics*, 5(1):011003–011003, 2018.