Representation, Alignment, and Generation: A Comprehensive Survey of Foundation Models for Non-Invasive Brain Decoding

Yifan Wang[a], Shaonan Wang[b], Yunhao Zhang[c,d], Changde Du[d,e], Cunhang Fan[f], Dongyang Li[g], Hongpeng Zhou[a], Hongyu Zhang[h], Jixing Li[i], Quanying Liu[g], Wei Huang[j], Yizhuo Lu[k,l], Zijiao Chen[m] and Jingyuan Sun[a,*]

[a]*Department of Computer Science, University of Manchester, Oxford Rd, Manchester, M13 9PL, Greater Manchester, United Kingdom*
[b]*Department of Language Science and Technology, The Hong Kong Polytechnic University, 8 Hung Lok Road, Hung Hom, Kowloon, Hong Kong*
[c]*State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[d]*School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China*
[e]*State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[f]*Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, Anhui, China*
[g]*Department of Biomedical Engineering, Southern University of Science and Technology, 1088 Xueyuan Avenue, Shenzhen, 518000, Guangdong, China*
[h]*School of Control Science and Engineering, Shandong University, Jinan, 250100, Shandong, China*
[i]*Department of Linguistics and Translation, City University of Hong Kong, Hong Kong*
[j]*School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, 610054, Sichuan, China*
[k]*Laboratory of Brain Atlas and Brain-Inspired Intelligence, State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[l]*School of Future Technology, University of Chinese Academy of Sciences, Beijing, China*
[m]*School of Humanities and Sciences, Stanford University, Stanford, 94305, CA, USA*

ARTICLE INFO

Keywords:
Brain-Computer Interface (BCI)
Non-Invasive Brain Decoding
Foundation Models
Deep Learning
Electroencephalography (EEG)
Functional Magnetic Resonance Imaging (fMRI)
Magnetoencephalography (MEG)
Functional Near-Infrared Spectroscopy (fNIRS)
Visual Decoding
Language Decoding
Speech Decoding
Auditory Attention Decoding (AAD)
Music Decoding

ABSTRACT

The ability to decode human thoughts, intentions, and perceptions directly from non-invasive brain recordings holds transformative potential for healthcare, communication, and human-computer interaction. However, translating the safety and scalability of methods like fMRI, EEG, and MEG into real-world utility has traditionally been hindered by low signal-to-noise ratios, limited spatial-temporal resolution, and the difficulty to collect large-scale high-quality data from an individual user. Recently, the emergence of Foundation Models (FMs)—large-scale, pre-trained artificial intelligence architectures—has catalyzed a paradigm shift in overcoming these obstacles. This survey provides a comprehensive overview of how FMs are redefining the boundaries of non-invasive brain decoding. We propose a unified methodological framework that synthesizes recent advancements into a coherent process: extracting robust, transferable representations from noisy neural signals; aligning these signals with the rich semantic spaces of pre-trained vision and language models; and leveraging powerful conditional generative priors to reconstruct high-fidelity outputs. We systematically review state-of-the-art applications across three key domains: visual reconstruction, language and speech decoding, and auditory processing. Furthermore, we critically examine the persisting challenges of computational efficiency, cross-subject generalization, and privacy governance. By mapping the current landscape and identifying key gaps, this work outlines a strategic research agenda aimed at transitioning FM-driven neurotechnology from laboratory proofs-of-concept to reliable, real-world applications.

*Corresponding author
yifan.wang-38@postgrad.manchester.ac.uk (Y. Wang); shaonan.wang@polyu.edu.hk (S. Wang); zhangyunhao2021@ia.ac.cn (Y. Zhang); changde.du@ia.ac.cn (C. Du); cunhang.fan@ahu.edu.cn (C. Fan); lidy2023@mail.sustech.edu.cn (D. Li); hongpeng.zhou@manchester.ac.uk (H. Zhou); hongyu.zhang@mail.sdu.edu.cn (H. Zhang); jixingli@cityu.edu.hk (J. Li); liuqy@sustech.edu.cn (Q. Liu); huangw@uestc.edu.cn (W. Huang); luyizhuo2023@ia.ac.cn (Y. Lu); zijiao@stanford.edu (Z. Chen); jingyuan.sun@manchester.ac.uk (J. Sun)
http://personalpages.manchester.ac.uk/staff/jingyuan.sun (J. Sun)
ORCID(s): 0009-0004-7864-6490 (Y. Wang); 0000-0002-1089-5472 (J. Sun)

1. Introduction

The enduring quest to understand the human brain and translate its intricate activity into meaningful outputs represents a central challenge in neuroscience and engineering [151]. Decoding brain signals—interpreting neural activity to infer cognitive states, intentions, or perceptual experiences—has garnered significant interest due to its profound potential. Laboratory research has yielded remarkable progress, demonstrating the feasibility of interpreting neural signals to control prosthetic limbs, enable communication for individuals with severe paralysis, and even reconstruct perceived visual stimuli from brain activity patterns [119, 83, 177]. These achievements underscore the transformative

possibilities, spanning healthcare innovations for neurological disorders, novel communication channels, and enhanced human-computer interaction paradigms [126].

To bring these technologies from the laboratory to the general public, researchers have increasingly focused on non-invasive brain decoding techniques, such as Electroencephalography (EEG), Magnetoencephalography (MEG), Functional Magnetic Resonance Imaging (fMRI), and Functional Near-Infrared Spectroscopy (fNIRS) [151, 13, 106, 48]. Unlike invasive methods, these techniques acquire neural signals from outside the body without surgical intervention, thereby avoiding associated health risks [141]. Their safety and scalability make them a promising foundation for widespread use. However, non-invasive methods face significant inherent challenges: they typically suffer from lower signal-to-noise ratios (SNR) and poorer spatial or temporal resolution compared to invasive implants. Furthermore, high-quality physiological data is relatively scarce, and data annotation is costly. These factors have significantly hindered the development of robust, generalizable decoding systems [2, 142].

Historically, brain decoding largely relied on linear models or task-specific deep networks. While effective in controlled paradigms, the performance of these models often collapses when applied to naturalistic environments or when generalizing across different subjects. In recent years, however, a paradigm shift in artificial intelligence (AI)—driven by the advent of Foundation Models (FMs)—has presented new opportunities to overcome these limitations [136, 59]. FMs are large-scale machine learning models pre-trained on vast and diverse datasets, often using self-supervised learning, which allows them to adapt to a wide range of downstream tasks [136, 31, 59]. Architectures such as the Transformer [31] have enabled models like GPT [137] in language and CLIP [136] in vision to capture intricate semantic patterns. By leveraging these capabilities for representation learning, feature extraction, and cross-modal alignment, FMs have significantly improved the fidelity and semantic accuracy of reconstructions from non-invasive brain recordings, opening unprecedented avenues for neurotechnology [59].

Although reviews on brain decoding are extensive, existing literature exhibits key shortcomings that this survey aims to address. Previous surveys have often been limited in scope or outdated relative to the rapid pace of AI development. For instance, reviews on EEG applications [170] or encoding/decoding [183, 2] often focus solely on specific modalities or emphasize algorithmic lists without unifying visual, linguistic, and multimodal fusion. Similarly, surveys on natural image reconstruction [138] and fMRI decoding [34] highlight limitations like data scarcity and hemodynamic delays but do not fully explore how modern FMs resolve these issues through "neuro-semantic alignment" and conditional generation. A unified framework that connects the three major task families—visual (image/video), language/speech, and cross-modal fusion—remains lacking.

This paper addresses this gap by proposing a comprehensive methodological framework: using transferable representations as a foundation to align brain signals with the semantic or acoustic latent space of large models. We review how "weak, sparse, and delayed" neural measurements can be transformed into readable images, text, and speech by performing conditional generation or retrieval within this aligned space. From this perspective, we systematically review advancements in image and video reconstruction, language and speech decoding, and auditory attention. We map datasets and evaluation metrics to this framework, emphasizing practical challenges such as cross-subject generalization, inference efficiency, and privacy governance.

The remainder of this survey is organized as follows. Section 2 delves into the specific characteristics and challenges of non-invasive neuroimaging modalities, using Figure 1 and Table 1 to illustrate data workflows and trade-offs. Section 3 introduces the fundamental concepts and architectures of FMs, summarizing relevant models in Table 3 to serve as a conceptual scaffold. Section 4 details state-of-the-art applications of FMs in decoding visual experiences, language, speech, and auditory information, supported by a navigation chart in Figure 2. In Section 5, we synthesize cross-domain trends and breakthroughs. Finally, Section 6 critically examines persisting challenges and outlines future research directions, and Section 7 concludes by summarizing the collaborative effort required to translate these technologies into real-world applications.

2. The Landscape of Non-Invasive Brain Decoding: Modalities and Challenges

To appreciate the potential impact of foundation models, it is essential to first understand the specific characteristics of non-invasive brain decoding techniques and the fundamental challenges they present (Summarized in Table 1).

2.1. Non-Invasive Neuroimaging Modalities

Non-invasive recordings in BCI and brain decoding are dominated by EEG, MEG, fMRI, and fNIRS, which differ in their physical principles and therefore in temporal/spatial resolution, robustness to noise and motion, cost, and deployability. These differences directly shape what can be decoded, how models should be designed, and how results are ultimately interpreted.

Electroencephalography (EEG) captures voltage fluctuations on the scalp generated by synchronous activity of large neuronal populations [151, 13]. Its chief advantage is millisecond-level temporal resolution, which enables tracking rapid neural dynamics and supports closed-loop or real-time applications [200]. EEG systems are relatively inexpensive, portable, and easy to deploy, contributing to their ubiquity in research and clinical settings. At the same time, volume conduction blurs spatial specificity, typically limiting spatial resolution to the centimeter scale [151]; recordings are also vulnerable to physiological artifacts (eye blinks, muscle activity) and environmental electrical noise [191].

Magnetoencephalography (MEG) measures the minute magnetic fields produced by the same neural currents that underlie EEG signals [151, 13]. Because magnetic fields are less distorted by skull and scalp than electric potentials, MEG generally offers better source localization (millimeters to centimeters) while preserving millisecond temporal precision [151]. These benefits come with practical constraints: MEG systems are expensive, require magnetically shielded environments, and are sensitive to head motion, which limits portability and broad accessibility.

In contrast to the electrophysiological modalities, **Functional Magnetic Resonance Imaging (fMRI)** infers neural activity indirectly via blood-oxygen-level-dependent (BOLD) signals [106]. The technique excels in spatial resolution at the millimeter scale, enabling precise mapping of cortical and subcortical representations and supporting decoding of fine-grained spatial patterns such as visual feature maps [106, 119, 83]. Its principal limitation is temporal: the hemodynamic response peaks several seconds after neural events, resulting in second-level temporal resolution [106]. Practical considerations—high scanner cost, immobility, acoustic noise, and strict motion constraints—further narrow its real-time applicability.

Functional Near-Infrared Spectroscopy (fNIRS), which estimates changes in oxy- and deoxy-hemoglobin using near-infrared light, shares the vascular origin of fMRI signals while offering greater portability and lower cost [48]. It provides an intermediate operating point with temporal resolution on the order of seconds and spatial resolution at the centimeter scale, and it is relatively tolerant to modest movement compared with EEG or fMRI [48]. However, limited penetration depth restricts sensitivity primarily to superficial cortical tissue, and spatial precision and coverage remain below those of fMRI [48].

Taken together, these modalities impose fundamental trade-offs (Table 1). High temporal resolution in EEG and MEG is advantageous for rapid interaction and control, but weaker spatial localization complicates decoding of fine spatial structure. fMRI provides the inverse profile—excellent spatial precision that supports detailed representational decoding, offset by poor temporal resolution that hinders fast feedback. fNIRS occupies a pragmatic middle ground, improving deployability while rarely achieving the best-in-class performance in either domain [48]. These properties dictate modeling priorities: EEG/MEG pipelines benefit from sophisticated temporal processing and artifact mitigation, whereas fMRI pipelines emphasize high-dimensional spatial feature extraction and alignment. Ultimately, the ceiling on any decoding approach—including those leveraging foundation models—is tightly coupled to the information content and quality of the chosen signal source [151, 106, 48]. Figure 1 summarizes the typical workflow for non-invasive brain data acquisition and preprocessing using fMRI, EEG, and MEG. This figure not only illustrates how raw measurements are transformed into analyzable time series and brain maps (such as BOLD time series, ERP/ERF), but also highlights the core preprocessing steps for different modalities. These processed outputs serve as direct inputs for downstream FM-based decoders during training and inference, thus supporting the discussion of the tradeoffs in Table 1 at the data level.

2.2. Fundamental Challenges in Non-Invasive Decoding

Beyond modality-specific limitations, non-invasive brain decoding faces a set of cross-cutting obstacles that fundamentally constrain robustness and scalability [2, 109, 55, 150]. First, signals are weak and easily corrupted: propagation through tissue attenuates and distorts neural activity, while eye blinks, muscle activity, and environmental interference introduce non-neural components that are difficult to remove without damaging informative structure [191, 55]. Second, the data are high-dimensional, non-stationary, and idiosyncratic—varying across and within individuals due to anatomy, physiology (fatigue, arousal), cognitive state (attention), and task demands—so models that perform well in one session or subject often fail to generalize without costly recalibration [109, 24]. Third, hard physical limits in spatial and temporal resolution cap extractable information: slow hemodynamics blur rapid events in fMRI [106], while volume conduction in EEG mixes sources and degrades spatial separability [151].

Practical deployment adds further constraints. Many applications require end-to-end latencies on the order of hundreds of milliseconds, forcing a difficult accuracy, latency, compute trade-off, particularly on portable or embedded hardware [202, 123]. Ecological validity is another barrier: models tuned on tightly controlled laboratory paradigms (e.g., P300 [131, 153], SSVEP [199], or motor imagery [6, 111]) often degrade in naturalistic settings with unconstrained behavior, rich sensory input, and movement, creating a persistent "lab-to-life" gap [109, 2]. Finally, data scarcity remains chronic: collecting large-scale, well-annotated non-invasive datasets is expensive and time-consuming, limiting the training of capacity-hungry models and amplifying risks of spurious correlations and distribution shift [76, 150].

These challenges jointly shape algorithmic priorities for non-invasive decoding: aggressive but principled denoising and artifact suppression; adaptation to non-stationarity and cross-subject variability; architectures that respect each modality's information bottlenecks; streaming inference with bounded latency; and training regimes that mitigate data scarcity (e.g., self-supervision, cross-subject/domain adaptation, and careful regularization). The attainable performance of foundation-model-based approaches is therefore tightly coupled to how effectively they confront these constraints.

2.3. Limitations of Pre-Foundation Model Approaches

Traditional computational methods often struggled to fully overcome these challenges. Early approaches using linear models (e.g., LDA, linear SVM, regression) [28, 179]

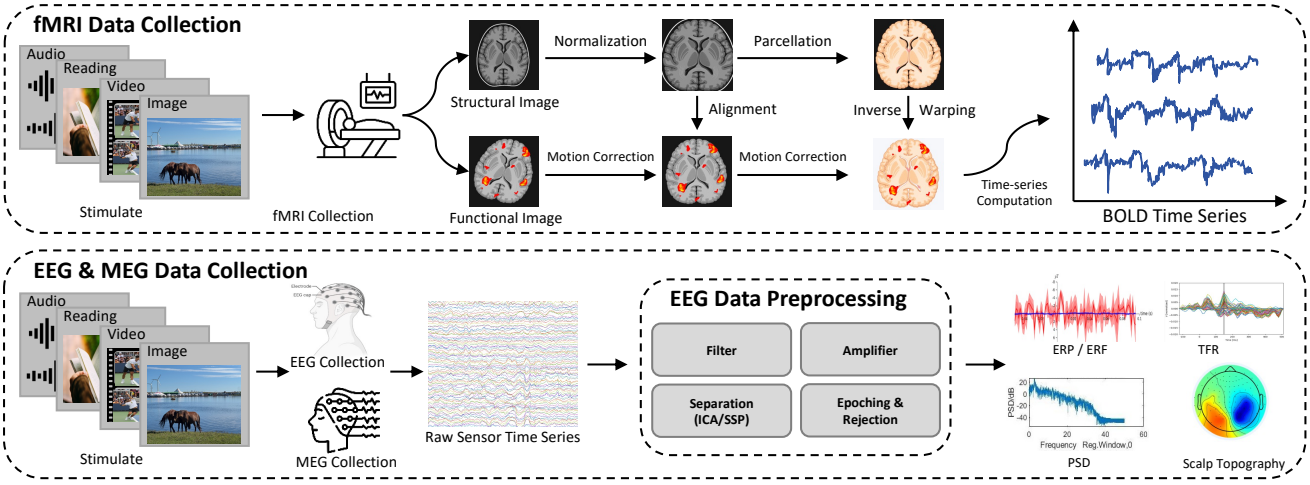

Figure 1: Non-invasive data collection and preprocessing across fMRI, EEG and MEG. The diagram highlights typical stimuli, acquisition hardware, core preprocessing steps (e.g., normalization/alignment/motion correction for fMRI; filtering/ICA/epoching for EEG/MEG), and the resulting analysis-ready outputs (BOLD time series, ERPs/TFR/PSD/topographies).

Table 1
Comparison of Non-Invasive BCI Modalities

Modality	Principle	Spatial Resolution	Temporal Resolution	Invasiveness	Portability	Cost	Key Strengths	Key Weaknesses
EEG	Electrical potentials from neural currents	Low (cm)	High (ms)	Non-invasive	High	Low	High temporal res., portable, low cost	Low spatial res., susceptible to artifacts/noise
MEG	Magnetic fields from neural currents	Moderate (mm-cm)	High (ms)	Non-invasive	Low (requires shielding)	Very High	High temporal res., better spatial res. than EEG	Expensive, requires shielded room, sensitive to movement
fMRI	Blood-Oxygen-Level-Dependent (BOLD) signal	High (mm)	Low (seconds)	Non-invasive	Very Low (immobile)	Very High	High spatial resolution	Low temporal res., indirect measure, expensive, immobile, noisy
fNIRS	Light absorption by blood hemoglobin	Moderate (cm)	Moderate (seconds)	Non-invasive	Moderate-High	Moderate	Portable, balance of resolution, movement tolerant	Lower spatial res. than fMRI, limited depth penetration

were efficient but lacked capacity for complex non-linear neural dynamics. For instance, linear models in Auditory Attention Decoding (AAD) (e.g., SR, TRF, CCA) [179, 49, 27, 32] showed declining performance in short time windows crucial for real-time use [15, 169].

Deep learning (CNNs, RNNs) offered improvements by automatically learning features and temporal dependencies [26, 198], enhancing tasks like motor imagery classification [6], ERP detection [153], fine-grained semantic decoding [173], and even early visual reconstruction [122, 11, 148]. However, these models, often trained from scratch on limited BCI datasets, still struggled with generalization due to data scarcity and variability [109], had difficulty decoding complex semantic content, and often required extensive preprocessing [26].

2.4. The Promise of Foundation Models for BCI Challenges

Foundation models (FMs) offer a principled route to address several long-standing hurdles in non-invasive decoding by leveraging large-scale pretraining and transfer learning. Trained on vast neuroimaging or multimodal corpora, FMs can internalize the statistical structure of brain signals and their relations to language and vision, thereby providing robust priors for downstream decoding under data and noise constraints.

First benefit concerns low SNR and artifacts. Self-supervised objectives—such as masked reconstruction in Masked Autoencoders (MAEs)—encourage models to capture invariant structure in unlabeled recordings, improving denoising and feature extraction without relying on brittle preprocessing [191, 64, 24]. Because these objectives learn to predict missing or corrupted segments from context, they naturally attenuate physiological and environmental noise while preserving informative components of the signal.

Second, scale and diversity in pretraining mitigate signal complexity and variability across subjects, sessions, and tasks. By exposing encoders to heterogeneous distributions, FMs learn more transferable representations that reduce subject-specific calibration and improve cross-dataset generalization; fine-tuning or lightweight adaptation then specializes the model to new cohorts with limited labels [109, 164]. This approach directly targets the non-stationarity that undermines conventional pipelines.

Third, although FMs cannot break the physical limits of spatial or temporal resolution, their capacity supports richer integration across sensors, space, and time—amortizing information over extended contexts to infer latent neural states that are only weakly expressed in raw measurements. In parallel, pretrained semantic spaces from language and vision (e.g., LLMs, CLIP) provide high-level targets that align with cortical representations [136, 116, 72, 59]. Mapping brain activity into these structured manifolds enables more expressive decoding—such as detailed image reconstructions or continuous speech content—than task-specific classifiers alone [145, 104, 164]. When coupled with conditional generators (diffusion models, autoregressive LLMs), strong priors can produce plausible outputs from noisy inputs, effectively regularizing the ill-posed inverse problem [162, 137, 59].

Finally, FMs help ameliorate data scarcity. Large-scale pretraining on web-scale text/images and emerging brain datasets bootstraps powerful initializations that require far fewer task-specific labels to reach competitive performance, and self-supervised learning on unlabeled neurodata further expands usable training signal [76, 59, 73]. In essence, by importing knowledge from broader corpora and aligning it with neural measurements, FMs offer a unifying recipe—robust representation learning, transferable adaptation, semantic alignment, and generative priors—that can accelerate progress toward reliable, non-invasive BCI systems.

2.5. Datasets & Benchmarks

In non-invasive brain decoding research, commonly used datasets can be systematically divided according to stimulus type, including image, video, speech/audio, reading/language, and multi-modal types. Different types of stimuli correspond to different neural representation characteristics and experimental paradigms, covering multiple levels from low-level vision to high-level semantic processing. Table 2 provides an overview of the datasets used in this article, along with their key elements (Dataset, Modality, Task/Stimuli, Standard BCI Paradigm, Research Paradigm, Subjects), to facilitate quick location and comparison.

2.5.1. Image Stimulus

Image datasets are fundamental to visual decoding research, primarily focusing on static scene and object representations. Typical datasets include NSD [5] and GOD [69], which are used for high-resolution brain imaging of natural scenes and category recognition, respectively; EEG-AVC [155] and Things-EEG2 [57] utilize high-temporal resolution EEG to record rapid visual categorization. Things-MEG [65] extends the same task to MEG, providing millisecond-level temporal dynamics. These datasets often employ the paradigm of subjects viewing natural or object images and are suitable for training and validating visual encoding and decoding models.

2.5.2. Video Stimulus

Video datasets emphasize temporal information and ecological validity. These include DNV [176], STN [146], and HAD [197] in fMRI, which are used for natural scene or action recognition tasks. In EEG, SEED-DV [102] and MOABB [146] combines rich semantic annotations with multimodal signals for semantic decoding from EEG to video. MEG Narrative [8] and HD-DOT [167] provide natural viewing data using high temporal resolution and high-density optical imaging. This type of video data supports temporal modeling, event boundary recognition, and representation learning of complex visual semantics.

2.5.3. Audio Stimulus

Auditory and speech datasets typically focus on natural narratives. For example, CLSR [164], NL fMRI [91], and Narratives fMRI [120] use fMRI to record brain activity while participants listen to natural stories. The Narratives fMRI dataset [120] is large (345 subjects), making it a landmark resource for natural language decoding. The EEG Imagined Speech Dataset (ISD) [182] covers both real and imagined speech tasks, supporting end-to-end research on direct speech generation from brain signals.

2.5.4. Reading/Language Stimulus

Reading and language datasets focus on the neural mechanisms of language processing. ZuCo [68] combines EEG and eye tracking to record brain activity while participants read natural sentences. MOUS [144] and Narratives fMRI [120] cover reading and auditory language tasks, and provide structural and diffusion MRI for multimodal fusion. ISD [182] involves speech imagery and language production, and is suitable for brain decoding tasks related to language production.

2.5.5. Multi-stimuli

Some datasets cover a variety of stimulus and task types, supporting multi-paradigm comparisons and cross-modal studies. For example, both HCP [36] and MOUS [144]

Table 2
Summary of datasets. Summary of datasets grouped by imaging modality and stimulus type used or referenced in this work. Columns list Modality (fMRI/EEG/MEG/HD-DOT), Stimulus (image, video, audio, reading, or multi-stimuli), Dataset (name), Task/Stimuli (experimental task and presentation format), Paradigm (commonly used cognitive-neuroscience/neuroimaging paradigms; standard BCI labels MI/ERP/SSVEP are used where applicable), and Subjects (number of participants).

Modality	Stimulus	Dataset	Task / Stimuli	Paradigm	Subjects
fMRI	Image	NSD [5]	Natural Scene Pictures	Natural Vision	8
		GOD [69]	Object Category Images	Category Identification / Representation	5
	Video	DNV [176]	Natural Video Viewing	Naturalistic Video	3
		STN [146]	Natural Video Viewing	Naturalistic Video	1
		HAD [197]	Human Action Recognition	Category Identification / Representation	30
	Audio	CLSR [164]	Naturalistic Narrative	Naturalistic Narrative	7
		NL fMRI [91]	Naturalistic Narrative	Naturalistic Narrative	8
		Narratives fMRI [120]	Naturalistic Narrative	Naturalistic Narrative	345
	Multi Stimuli	HCP [36]	Multi Task	Multi-paradigm Resource	1200
	Reading/Audio	MOUS [144]	Language Processing	Multi-paradigm Resource	204
EEG	Image	EEG-AVC [155]	Visual Object Classification	ERP	6
		Things-EEG2 [57]	Fast Sequence of Natural Images	ERP	10
	Video	SEED-DV [102]	Natural Video Viewing	Naturalistic Video	20
		MOABB [146]	Multi Task	MI / ERP / SSVEP	-
	Reading	ZuCo [68]	Reading / Language Comprehension	ERP	12
	Reading/Imaging Reading	ISD [182]	Imagine Pronunciation	Speech Imagery	13
MEG	Image	Things-MEG [65]	Natural Scene Pictures	Category Identification / Representation	4
	Multi Stimuli	HCP [36]	Multi Task	Multi-paradigm Resource	95
	Reading/Audio	MOUS [144]	Language Processing	Multi-paradigm Resource	204
	Video	MEG Narrative [8]	Naturalistic Narrative	Naturalistic Narrative	3
HD-DOT	Video	HD-DOT [167]	Natural Video Viewing	Naturalistic Video	3

offer large-scale structural and functional MRI (HCP includes 1,200 subjects), covering multiple tasks such as reading, hearing, and vision, and providing additional information such as genetics and behavioral data. MOABB [146] integrates multiple public EEG datasets and standardizes paradigms such as MI, ERP, and SSVEP, making it a valuable resource for algorithm comparison and benchmarking.

3. Foundation Models: Core Concepts and Architectures for Brain Decoding

Understanding the fundamental principles and specific architectures of FMs is crucial for appreciating their application in non-invasive brain decoding.

3.1. Defining Foundation Models

Foundation Models (FMs) refer to a class of large-scale artificial intelligence models that are trained on massive and diverse datasets—often using self-supervised learning objectives—and subsequently adapted to a wide range of downstream tasks through fine-tuning or prompting strategies [136, 31]. Unlike traditional task-specific models, FMs acquire general-purpose representations that can be reused across domains and modalities, making them particularly powerful in settings where labeled data is scarce or heterogeneity is high.

At their core, FMs combine scale with broad pre-training. Parameter counts span from hundreds of millions to trillions, and training data frequently reach web or multi-modal scale; such regimes enable models to internalize high-order statistical regularities and support emergent capabilities beyond those of smaller counterparts [81, 66, 67]. Pre-training objectives instantiate this breadth: masked-token prediction in language (e.g., BERT) and cross-modal contrastive alignment in vision–language (e.g., CLIP) endow models with semantic structure, syntactic competence, and cross-modal correspondences without manual annotation [31, 136].

Equally central is adaptability. Because pre-training yields versatile representations, modest amounts of task-specific data often suffice to specialize an FM for classification, generation, retrieval, or reasoning under few-/zero-shot settings. In neuroimaging, this adaptability directly addresses chronic data scarcity and subject variability by enabling lightweight transfer to new cohorts, sessions, or paradigms [164].

Prominent exemplars include large language models (GPT [137], BERT [31]), vision–language encoders (CLIP [136]), and modern generative models (e.g., diffusion backbones used for image synthesis and reconstruction [162]). Conceptually, the paradigm shifts from crafting bespoke pipelines to leveraging a powerful, pre-trained base that can

Table 3
Overview of Key Foundation Model Architectures applied in BCI

Modality	Method	Architecture	Pretraining Task
Language	BERT	Transformer (Encoder-only)	Masked Language Modeling (MLM), Next Sentence Prediction (NSP)
	GPT Series	Transformer (Decoder-only)	Autoregressive (Next Token Prediction)
	LLaMA	Transformer (Decoder-only)	Autoregressive (Next Token Prediction)
Vision	ViT	Transformer (Encoder-only)	Supervised or Self-supervised
	Stable Diffusion	U-Net	Denoising Diffusion
	StyleGAN	GAN	Adversarial Learning
Speech	Wav2Vec 2.0	CNN + Transformer	Masked Prediction on Latent Speech Representations
	Contrastive EEG-Speech	Hybrid Dual-Encoder	Contrastive Learning
Neuroimaging	M-DBPNet	CNN	Supervised/Feature Extraction
	ListenNet	CNN-RNN	Supervised
	MHANet	Transformer (Encoder-only)	Supervised
	EEG-Graph Nets	GNN	Graph-based Modeling
	DGSD	GNN / Graph U-Net	Graph-based Modeling
	MAE for fMRI/EEG	Transformer (Encoder-only)	Masked Autoencoding
Language & Vision	CLIP	Dual Encoder (Text & Image)	Contrastive Learning
	BLIP	Transformer (Encoder-only)	Image–Text Contrastive, Image–Text Matching, Captioning

be aligned with specific BCI objectives through fine-tuning or prompting.

3.2. Key Properties Driving BCI Advancement

Foundation models advance non-invasive decoding chiefly through three mutually reinforcing properties. First, they learn high-quality, hierarchical representations that align with neural population codes. Vision and language FMs capture semantic structure across layers [136, 137]; these internal embeddings correlate with brain activity during corresponding tasks—e.g., fMRI responses in visual cortex track layers of vision encoders and vision–language models such as CLIP, whose embedding space aligns with neural responses to images and text [116, 136, 104, 59]. Beyond holistic embeddings, recent work demonstrates that disentangling conceptual representations into fine-grained semantic subdimensions can significantly enhance the interpretability and fidelity of mapping neural activity to semantic meaning [194]. Notably, larger language models often exhibit stronger brain alignment, suggesting scaling benefits for representational fidelity [7, 53]. In addition to unsupervised language models based on English, research on Chinese has also shown that efficient parameter tuning methods can significantly affect language-brain alignment [194].

Second, broad pre-training underwrites transfer and generalization. Because FMs are exposed to heterogeneous, large-scale corpora, their features transfer with modest adaptation to small BCI datasets, improving robustness across subjects, sessions, and paradigms and mitigating the chronic data-scarcity problem [109, 164, 93]. This adaptability enables few-/zero-shot behaviors that are difficult to obtain with narrowly trained decoders.

Third, strong conditional generation closes the loop from brain signals to rich outputs. Diffusion [162] backbones and autoregressive LLMs [137] can be conditioned on FM-derived neural embeddings to synthesize images or text, using powerful priors to regularize the ill-posed inverse problem and produce plausible reconstructions even under noisy conditioning [25, 158, 24, 145]. Together, neural alignment, transferable representations, and generative priors create an end-to-end pathway from weak non-invasive signals to high-level percepts and concepts, thereby accelerating progress in reliable BCI decoding [59].

3.3. Relevant Architectures and Their Roles in BCI

Neuroimaging-native architectures exploit the structure of brain data itself. Purpose-built encoders (e.g., M-DBPNet, ListenNet, MHANet) tailor feature extraction to fMRI/EEG characteristics, while graph-based models (e.g., EEG graph networks, DGSD) encode inter-sensor topology and functional connectivity to better capture mesoscale dynamics [39, 37, 96, 38]. Together, these modality-specific roles illustrate how FMs function not as a monolith but as a toolbox: language decoders contribute generative priors, vision encoders and contrastive bridges supply structured semantic targets, speech encoders provide acoustically grounded manifolds, and neuroimaging models respect the physics and

topology of the measurements—collectively advancing the fidelity and robustness of non-invasive decoding.

Recent advances in non-invasive brain decoding have been propelled by the targeted adoption of diverse foundation-model (FM) architectures across language, vision, speech, and neuroimaging, each filling a distinct role rather than occupying the same rung in a uniform taxonomy (Table 3). In language, transformer families split naturally into encoder-only and decoder-only uses. Encoder-only models such as BERT provide context-rich embeddings via masked-token objectives that serve as stable targets or intermediate representations for textual decoding and semantic alignment [31]. Decoder-only models—including GPT and LLaMA—offer strong autoregressive priors that support coherent sequence generation from neural features, enabling reconstruction or continuation of linguistic content under noisy or partial conditioning [137, 166].

In vision, encoder-centric ViTs and masked-autoencoding regimes (e.g., MAE) learn robust patch-level structure that is well matched to spatially distributed neural signals, especially for fMRI-based decoding where fine-grained representational geometry matters [33, 64]. Dual-encoder models such as CLIP align images and text through contrastive learning and thus provide a shared semantic manifold that links brain activity to multimodal concepts, improving both zero-/few-shot decoding and retrieval [136]. For generative targets, diffusion backbones now dominate high-fidelity reconstructions from brain activity, superseding earlier GAN-based [61] approaches for naturalistic images by leveraging powerful priors and iterative denoising.

For speech, self-supervised audio encoders like Wav2Vec 2.0 capture latent phonetic and prosodic structure through masked prediction on large unlabeled corpora, yielding representations that can be aligned to EEG/MEG correlates of auditory perception or imagery [9]. Complementary contrastive dual-encoder formulations map speech and neural activity into a shared space, facilitating recognition or retrieval of perceived/imagined speech with minimal labels [135].

4. Foundation Models Driving Advancements in Non-Invasive Brain Decoding

The integration of FMs has catalyzed significant progress, achieving unprecedented accuracy and richness in interpreting brain signals related to vision, language, speech, and auditory attention. A common conceptual pipeline involves: 1) Extracting denoised representations from neural signals (often using MAEs or SSL [191, 24, 76, 73]), 2) Connecting neural signals with decoding targets by aligning neural representations with FM semantic spaces (e.g., CLIP, LLM embeddings) using mapping modules and contrastive learning [104, 164, 93, 23], and 3) Generating decoding results using conditional FMs (Diffusion Models, LLMs) guided by the aligned neural representations, leveraging strong priors for high-quality output [25, 158, 24, 145]. To provide a context for recent research, Figure 2 maps representative works from 2017–2025 across modalities and output types; we use it as a roadmap for the subsections that follow.

4.1. Foundation Models in Neuroimaging

In non-invasive brain decoding, foundation models (FMs) typically play three complementary roles: First, they utilize self-supervised/weakly supervised methods to obtain transferable representations, such as MAE [64], wav2vec 2.0 [9], and BLIP/CLIP [95], providing a robust foundation for small sample sizes and cross-subject learning [136, 140]. Second, they map fMRI/EEG/MEG to the semantic or acoustic manifolds of pre-trained models through neuro-semantic/acoustic alignment. For example, aligning brain representations to the visual-linguistic space of CLIP/BLIP or the acoustic space of wav2vec 2.0 [104, 136, 140]. Third, they leverage strong generative priors (diffusion/autoregression) to transform weak and time-delayed neural conditions into readable image, text, and speech outputs [140, 162, 125, 108]. A representative approach in recent years is "fMRI → CLIP semantic embedding → latent space diffusion reconstruction," which significantly outperforms earlier task-specific networks in terms of semantic consistency and perceptual quality [164, 104]. Meanwhile, native neuroimaging models (such as M-DBPNet [39], ListenNet [37], MHANet [96], EEG graph networks/self-distilled graph networks, and MAE for fMRI/EEG) explicitly encode sensor topology and measurement physics to compensate for the modal gap in general FMs. Overall, the triadic collaboration of transferable representation + neural-semantic alignment + generative priors is connecting non-invasive weak signals end-to-end to high-level readable content, demonstrating stable gains in small-sample, cross-subject, and zero/small-sample scenarios.

At the same time, rapidly evolving role for foundation models focuses on learning universal representations of neuro-neuronal connectivity. These models, often termed Neuro-Connectome Transformers or Graph Foundation Models, directly consume fMRI/EEG time-series or derived functional connectivity matrices. Instead of learning semantic relationships like LLMs, their primary objective is to develop generalized feature extractors that capture the complex, dynamic linking mechanisms and rules between distinct neural populations [147, 85]. This approach often leverages Graph Transformer architectures to model the brain as a dynamic graph, employing self-attention mechanisms to discover high-order, global dependencies across brain regions, leading to powerful, transferable representations for tasks like brain disorder diagnosis and biotyping, even across large-scale heterogeneous datasets [175]. Furthermore, techniques like Masked Brain Modeling (MBM), applied directly to raw fMRI data, serve as a potent form of self-supervised learning to force the model to infer and thus implicitly encode the underlying functional relationships between masked and unmasked brain areas [24]. This paradigm shift moves neuroimaging analysis from fixed,

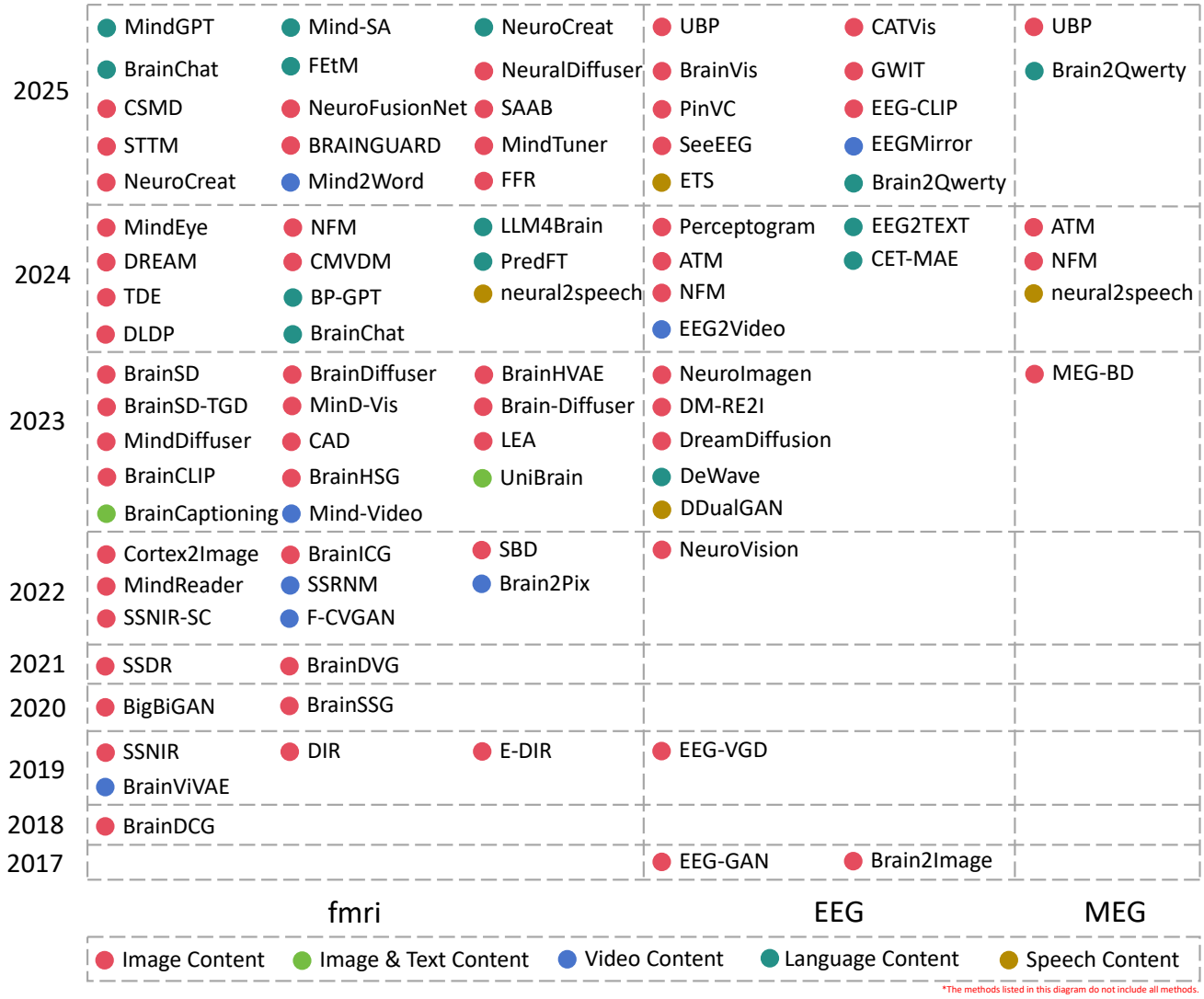

Figure 2: Chronological map of representative non-invasive brain-decoding works (2017–2025) across modalities (fMRI/EEG/MEG). Colors denote output content types (image, video, language, speech). This overview situates subsequent method sections and helps cross-reference modality-specific trends.

statistical connectivity measures towards flexible, learnable, and generalizable neuro-relational embeddings.

4.2. Decoding Visual Experiences

Visual decoding seeks to reconstruct perceived images or videos from brain activity by coupling neural embeddings with powerful priors from foundation models. Between 2023 and 2025, progress has accelerated as CLIP-style semantic spaces and diffusion backbones became standard components across modalities (overview in Table 4; schematic in Figure 3). In fMRI, spatial resolution is key [101]. The dominant strategy combines CLIP's representational power with diffusion models' generative prowess [59]. An encoder maps fMRI signals to CLIP's shared embedding space, capturing semantics [59, 104]. These embeddings condition a latent diffusion model (e.g., Stable Diffusion [162], Versatile Diffusion [123]) to generate the image [145, 59, 123, 159]. CLIP provides semantic guidance, while diffusion models ensure fidelity [24, 59]. Foundational work like MinD-Vis (CVPR 2023) used sparse masked modeling (MAE-like) for fMRI representation and a double-conditioned LDM (DC-LDM) for reconstruction, addressing variability [24]. Brain-Diffuser predicted multimodal CLIP features from fMRI to condition Versatile Diffusion [123]. These FM-based methods yield significant improvements in semantic accuracy and perceptual fidelity over earlier approaches [24, 101].

With EEG, the emphasis shifts to exploiting millisecond dynamics under limited spatial precision. Recent end-to-end schemes learn a projection from EEG to CLIP space—e.g., via an Adaptive Thinking Mapper—and then invoke diffusion priors guided by the CLIP embedding, a coarse EEG-derived image, and auto-captions to stabilize generation, achieving state-of-the-art zero-/few-shot performance

and indicating that informative semantic structure emerges around 250 ms post-stimulus [93, 26, 10, 154]. MEG follows a similar strategy: neural activity is aligned to pretrained visual embeddings (CLIP, DINOv2) through regression or contrastive learning, and a diffusion model is subsequently conditioned to reconstruct images; time-resolved analyses reveal peaks that track high-level semantics and outperform linear decoders by large margins in retrieval settings [12, 123].

Extending from static images to video requires addressing the temporal mismatch between second-scale hemodynamics and high frame rates. Recent fMRI systems therefore separate *semantics* from *perception*: windowed BOLD signals are mapped to CLIP-like keyframe embeddings while a complementary pathway estimates low-level appearance and motion flow; both streams condition a pretrained text-to-video diffusion model to yield temporally coherent clips with improved SSIM and cross-frame consistency [59, 25, 201, 157].

Overall, fMRI remains strongest for spatial detail, whereas EEG and MEG excel at temporal dynamics; FM-based alignment (e.g., CLIP) and generative priors (diffusion) compensate for each modality's weaknesses and can be adapted with minimal architectural changes across recording types [101, 26, 123, 93]. This convergence on "neural-to-semantic alignment + conditional diffusion generation" constitutes a unifying blueprint for reconstructing naturalistic visual experiences from non-invasive signals.

4.3. Decoding Language and Speech

Recent progress in non-invasive language and speech decoding has come from treating foundation models (FMs)—large language models (LLMs) and self-supervised sequence/speech encoders—as priors, targets, and generators within end-to-end pipelines (overview in Table 5; schematic in Figure 4) [143, 23, 135]. The central idea is to align neural activity to pretrained semantic or acoustic manifolds and then exploit powerful generative or retrieval mechanisms to recover continuous content.

Despite fMRI's slow temporal dynamics, LLM-centric pipelines enable continuous semantic decoding by using the language model as a prior/decoder to enforce coherence [143]. A common recipe builds an encoding model that predicts fMRI responses from linguistic features [161] and then scores candidate continuations sampled from an LLM (e.g., GPT), selecting the sequence that best matches the predicted brain activity—trading data for fluency and substantially mitigating temporal limits [164]. Recent research has further scrutinized this alignment, revealing that while neural language models are powerful, psychologically plausible models can sometimes offer superior alignment with brain representations in specific linguistic contexts [193]. Performance scales with model size but typically requires sizable per-subject data; nonetheless, systems can recover the gist of perceived/imagined narratives or silent videos, indicating access to higher-level semantics beyond low-level cues [164, 7, 4]. For example: *MindLLM* targets subject-agnostic decoding via a neuroscience-informed attention encoder plus "Brain Instruction Tuning" with an off-the-shelf LLM [134]; *UMBRAE* aligns fMRI into a VLM latent space to flexibly decode text and/or images [181]; and *GPT-2 prompting* directly treats fMRI as prompts aligned to text prompts through contrastive learning [23]. Early language decoding work primarily focused on lexical-level decoding. Subsequently, research began to shift towards more complex sentence-level semantic representations [160]. Overall, FMs supply powerful linguistic constraints and generation, advancing versatility and cross-subject generalization while compensating for fMRI's temporal limitations [143, 164, 134].

EEG's millisecond precision shifts the focus to acoustic dynamics. Pipelines map EEG into pretrained speech representation spaces (e.g., Wav2Vec 2.0) using regression or contrastive objectives, after which CTC/sequence decoders or retrieval heads infer phonemes and words; this transfer learning improves robustness with limited labels and supports cross-subject generalization [26, 135]. The approach leverages FM encoders as invariant acoustic priors while allowing lightweight adaptation to individual subjects and sessions.

MEG occupies a complementary sweet spot: its temporal fidelity enables fine-grained analyses while offering better source separability than EEG. Neuroscience-inspired self-supervised learning on unlabeled MEG (masking, band prediction) yields representations that scale with data and generalize across tasks and subjects, setting strong baselines for phoneme/word decoding under data scarcity [73]. Aligning MEG features to pretrained speech models further boosts recognition of perceived speech and supports time-resolved probing of linguistic feature dynamics [73, 135].

Looking toward practical communication BCIs, non-invasive routes based on fMRI [164], potentially fNIRS [4], and EEG spellers augmented with FM priors [17] show growing promise but face constraints in equipment bulk, calibration demands, latency, and robustness. Across modalities, however, the FM blueprint—neural-to-semantic/acoustic alignment plus generative or retrieval decoding—provides a coherent path to richer language and speech reconstruction from weak non-invasive signals.

4.4. Decoding Audio, Music, and Auditory Attention (AAD)

Decoding auditory experiences—ranging from attended speech to complex music—has benefited from foundation-model (FM) ideas, though progress lags behind vision and language [128]. A central thread is auditory attention decoding (AAD), which aims to infer the attended source (e.g., speaker) from non-invasive brain signals to enable neuro-steered hearing aids [55]. Early linear pipelines (stimulus/temporal response functions, CCA) [28] provided a principled starting point [179, 49], but they struggle with nonlinearity and short decision windows that are critical for real-time control [132, 15, 169]. Recent systems

Table 4
Summary of Decoding Visual Experiences (Images and Video). Columns denote: Brain Modality (neuroimaging method), Decoded Content (target output such as image or video), Dataset (benchmark used), Representation Model (feature extraction model), Generation Model (framework for reconstruction), Work (Reference) (method name and citation), Method (Follows our taxonomy (DPC/FMA/E2E/AG-RAG/AG-EML)).

Brain Modality	Decoded Content	Dataset	Representation Model	Generation Model	Work (Reference)	Method
fMRI	Image	NSD [5]	CLIP	SDM	BrainSD [162]	DPC
			CLIP	SDM	BrainSD-TGD [163]	DPC
			VQ-VAE	SDM	MindDiffuser [108]	E2E
			CLIP	-	BrainCLIP [104]	FMA
			CLIP	LDM	BrainDiffuser [125]	DPC
			CLIP	DALL-E2	MindEye [145]	FMA
			CLIP	GAN	MindReader [98]	AG-RAG
			CLIP	LDM	DLDP [44]	AG-RAG
			CLIP	-	NeuroFusionNet [89]	DPC
			CLIP	DM	NeuralDiffuser [94]	FMA
			CLIP	DM	BRAINGUARD [165]	FMA
			CLIP	DM	MindTuner [60]	FMA
			CLIP	DM	CSMD [184]	FMA
			CLIP	DM	FFR [188]	FMA
			CLIP	-	NFM [45]	FMA
		GOD [69]	-	DM	CMVDM [189]	AG-RAG
			SC-MBM	DC-LDM	MinD-Vis [24]	E2E
			DC-MAE	LDM	CAD [158]	DPC
			CNN	LDM	SBD [46]	AG-RAG
			D-VAE	GAN	BrainDVG [139]	DPC
			-	GAN	BigBiGAN [117]	DPC
			SwAV ResNet-50	IC-GAN	BrainICG [124]	DPC
			VDVAE	LDM	Brain-Diffuser [114]	DPC
			CLIP	MaskGIT	LEA [133]	FMA
			CLIP	DM	SAAB [80]	FMA
		NSD&GOD [69, 5]	CLIP	SDXL-Turbo	NeuroCreat [78]	FMA
			CLIP	DM	STTM [105]	FMA
		ImageNet [30]	VGG19	LDM	[80]	DPC
	Image & Text	NSD [5]	CLIP	LDM	UniBrain [110]	DPC
			CLIP	DM	Brain Captioning [47]	DPC
	Video	DNV [176]	-	SDM	Mind-Video [25]	AG-RAG
			MARS	De-CNN	SSRNM [87]	DPC
			VAE(Encoder)	VAE(Decoder)	BrainViVAE [63]	DPC
			-	fMRI-GAN	f-CVGAN [171]	DPC
			CLIP	DM	Mind2Word [75]	FMA
		STN [146]	-	FCN	Brain2Pix [90]	DPC
	3D	Self Dataset [54]	CLIP	DM	MinD-3D++ [54]	FMA
EEG	Image	EEG-AVC [155]	CLIP	DM	NeuroImagen [88]	DPC
			EVRNet	DM	DM-RE2I [190]	E2E
			Visual-Guided CNN	GAN	EEG-VGD [77]	DPC
			-	cGAN	EEG-GAN [127]	AG-RAG
			-	GAN	NeuroVision [84]	DPC
			CLIP	DM	BrainVis [51]	FMA
			-	DM	GWIT [107]	FMA
		Things-EEG2 [57]	CLIP	LDM	Perceptogram [41]	DPC
			CLIP	DM	ATM [93]	E2E
			CLIP	-	UBP [180]	FMA
			CLIP	DM	PinVC [79]	FMA
			CLIP	SDXL-Turbo	EEG-CLIP [16]	FMA
			CLIP	DM	SeeEEG [86]	FMA
		Self-collected Dataset [62]	LSTM Network	DCGAN	Brain2Image [82]	DPC
		MOABB [74]	CLIP	DM	DreamDiffusion [10]	E2E
		ImageNet [30]	CNN	LDM	[43]	DPC
			CLIP	DM	CATVis [113]	FMA
		NSD [5]	CLIP	-	NFM [45]	FMA
	Video	SEED-DV [102]	CLIP	DM	EEGMirror [103]	FMA
MEG	Image	Things-MEG [65]	DINOv2, CLIP	LDM	MEG-BD [12]	FMA
			CLIP	-	UBP [180]	FMA
		Things-EEG2 [57]	CLIP	DM	ATM [93]	E2E
		NSD [5]	CLIP	-	NFM [45]	FMA

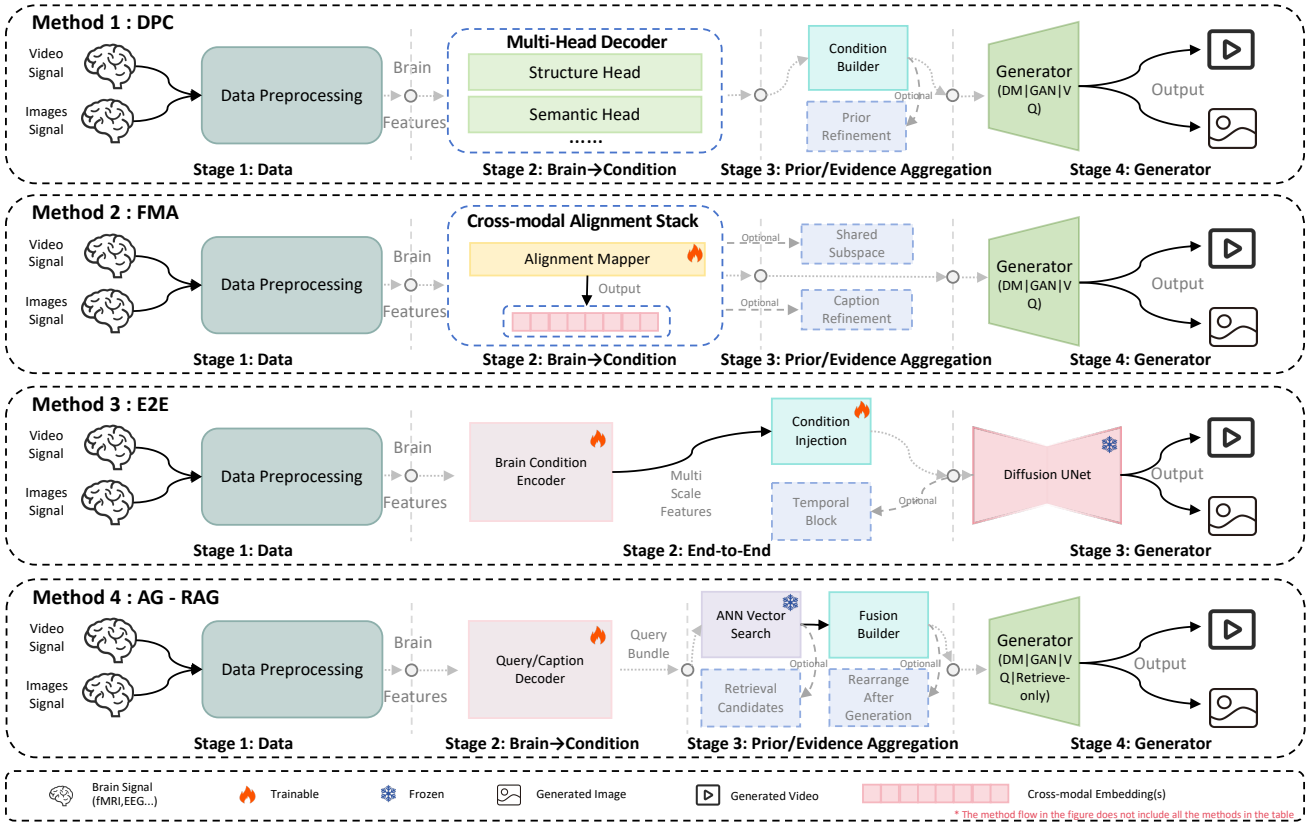

Figure 3: Decoding Visual Experiences Flowchart. This flowchart shows the different types of methods for images and video decoding tasks. (1)DPC (Decoded-Prior Conditioning): Multi-head decoders predict structural (VAE latent, depth/edges) and semantic (CLIP/caption) priors, fused and fed to a generator (DM/GAN/VQ); dashed boxes denote prior refinement/re-ranking options. (2) FMA (Foundation Model Adaptation (Alignment)): Brain features are aligned to multimodal embeddings (CLIP/text), optionally with shared-subspace learning and caption refinement, then passed to a generator (DM/GAN/VQ) and diffusion. (3) E2E (End-to-End Decoding): A learned brain condition encoder injects multi-scale features into a largely frozen diffusion UNet; an optional temporal block supports video. (4)AG-RAG (Augmented Generation via Retrieval-Augmented Generation): A query/caption decoder drives ANN search over a multimodal index; retrieved candidates supply latent init, geometric controls, and prompts to the generator (DM/GAN/VQ/Retrieve-only), with optional post-generation re-ranking.

replace these with non-linear encoders that borrow FM-inspired components—self-attention, convolutional hierarchies [38, 156, 185, 3], and graph neural networks—to model spatio-temporal structure and inter-sensor topology, yielding markedly higher accuracy at short latencies [14, 39, 37, 96]. Within this family, some approaches perform identity-based AAD by correlating EEG with reference speech streams using CNN/attention encoders or encoder–decoders [156, 185, 3]; others pursue reference-free spatial AAD by estimating the attended location directly from EEG via spatio-temporal attention, EEG-graph networks, dynamic graph self-distillation networks, or recent hybrid/mamba-style architectures [3, 38, 196, 121, 39, 186, 37]. Collectively, these non-linear models narrow the gap to practical neuro-steered hearing aids by improving robustness in short windows [55, 192].

Beyond attention, reconstruction of auditory content is emerging. For fMRI, hierarchical feature decoding leverages knowledge of the auditory pathway: models first decode deep-network–inspired auditory features from BOLD and only then synthesize waveforms using an audio generative model, thereby avoiding a direct slow-fMRI-to-fast-audio mapping [128]. For EEG, millisecond temporal precision suits rhythm and envelope tracking; recent work employs latent diffusion models for naturalistic music reconstruction, pushing beyond classical linear envelope decoding though typical quality remains below speech reconstruction [130, 152]. Notably, DMF2Mel [40] has achieved precise reconstruction of minute-level continuous imagined speech, effectively balancing temporal dependency modeling and information retention efficiency in long-sequence decoding. These results underline both the promise of generative priors and the challenge of capturing rich musical structure from sparse, noisy measurements.

Comparative analyses highlight systematic differences between music and speech decoding. Linear EEG models reveal shared basic acoustic tracking yet lower reconstruction quality for music envelopes, consistent with distinct

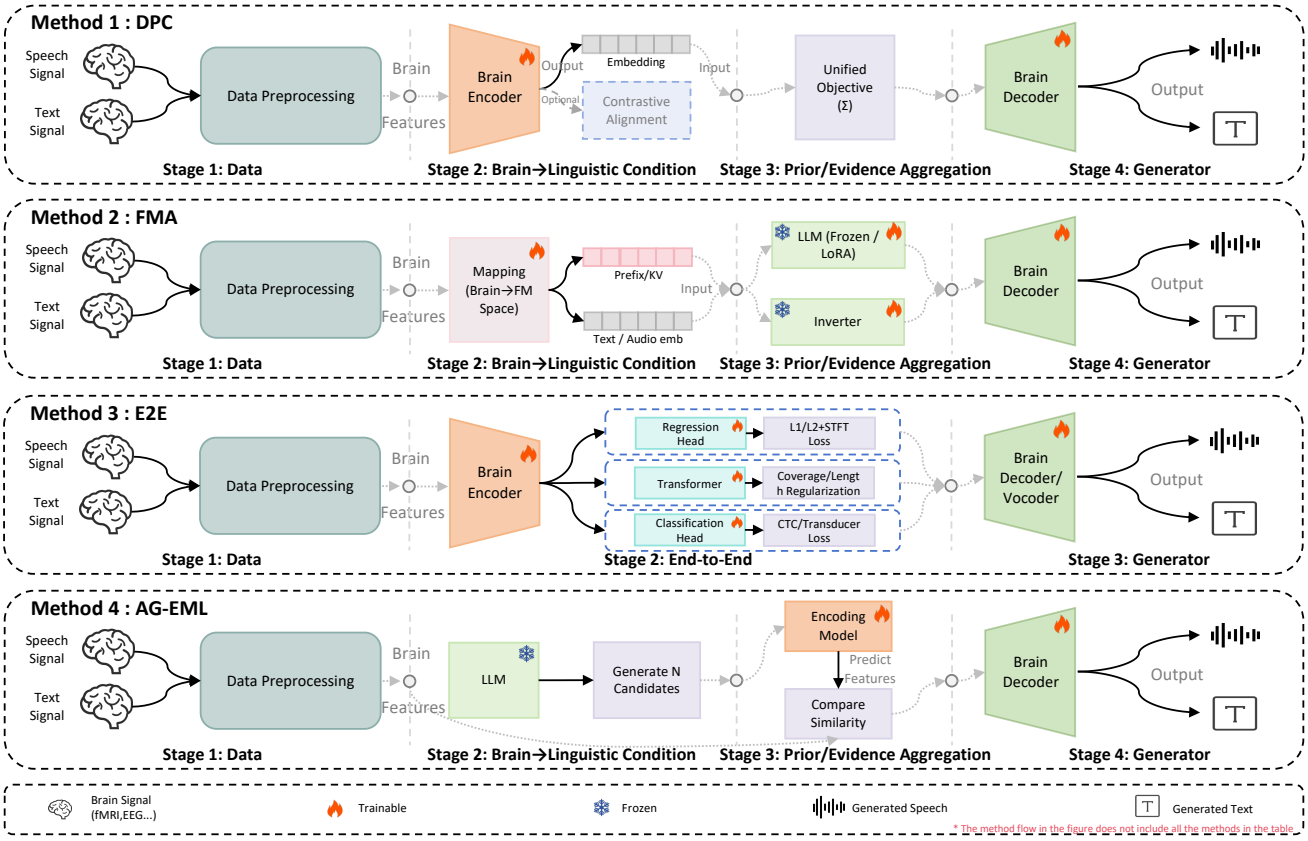

Figure 4: Decoding Language and Speech Flowchart. This flowchart shows the different types of methods for language and speech decoding tasks. (1) DPC (Decoded-Prior Conditioning): Preprocessed brain signals are encoded by a Brain Encoder and decoded by a Unified Decoder (text/speech) under a Regularized Multitask Objective (CE, spectral losses, alignment, regularization); the dashed Contrastive Alignment denotes an optional brain–semantic alignment. (2) FMA (Foundation Model Adaptation): A Mapping (Brain→FM Space) projects brain features into the foundation-model space, followed by two routes: (i) injecting context vectors/keys/prefixes to a frozen or lightly fine-tuned LLM; (ii) regressing to text/audio embeddings and using an Inverter to recover readable text or synthesizable speech. (3) E2E (End-to-End Decoding): After a shared Brain Encoder, three branches are used: Regression Head to acoustic features with L1/L2 + STFT and a Vocoder; Transformer seq2seq with coverage/length regularization; and Classification Head with CTC/Transducer loss for direct character/phoneme/word decoding. (4) AG-EML (Augmented Generation via Encoding-Model-in-the-Loop): A separately trained encoding model scores LLM-generated candidates against measured brain activity (NeuroScore), enabling reranking to select the final output.

cortical processing pathways and with the added structural complexity of music [152]. This complexity is being used as a probe for auditory scene analysis, while fMRI studies map genre-level representations across cortex [118]. Overall, FM-inspired encoders and generative decoders provide a coherent route for AAD and auditory reconstruction: learn invariant, topology-aware neural representations; align them with pretrained acoustic manifolds; and leverage strong priors for synthesis—steps that collectively improve short-window accuracy for attention control and expand the scope of non-invasive reconstruction of sound and music.

4.5. Multimodal Decoding

A growing line of work moves beyond single-modality pipelines to fuse complementary signals—across neural modalities (e.g., fMRI+MEG/EEG, EEG+fNIRS) and with auxiliary behavioral streams (e.g., eye tracking)—while aligning all channels to shared FM latents (CLIP/VLM or speech encoders) for generation or retrieval. Methodologically, these systems pursue two converging threads: (i) neural–neural fusion to combine fMRI's spatial precision with MEG/EEG's millisecond dynamics or fNIRS's portability, and (ii) neural–semantic alignment that maps each input stream into a unified FM-conditioned space where decoding and cross-modal transfer are performed with minimal architectural changes.

On the neural–neural side, unified frameworks such as UMBRAE learn a shared latent in which different brain measurements (and even different stimulus types) can be decoded to image or text outputs; contrastive and generative objectives encourage instance-level alignment and enable flexible "modality in, task out" decoding [181]. Complementary work explicitly aligns EEG, MEG, and fMRI representations in a common space to support decoding,

Table 5
Summary Table of Recent Language/Speech Decoding Works

Brain Modality	Decoded Content	Dataset	Representation Model	Generation Model	Work (Reference)	Method
fMRI	Text	Narratives fMRI [120]	-	Transformer (Decoder)	PredFT [187]	E2E
		NSD [5]	CoCa	Transformer	BrainChat [71]	FMA
			CLIP	GPT-2	MindGPT [22]	FMA
			CLIP	GPT-2	Mind-SA [21]	FMA
			CLIP	LLaMA-3	NeuroCreat [78]	FMA
	Narrative Semantics	CLSR [164]	-	GPT-1	[164]	AG-EML
	Auditory Text	NL fMRI [91]	-	GPT-2	BP-GPT [23]	AG-EML
	Image Captions	COCO [99]	DINOv2	GPT-2	[149]	FMA
	Image Captions	NSD [5]	CoCa	-	BrainChat [70]	FMA
	Descriptions of Images	NSD [5]	-	Vec2Text	From Eye to Mind [42]	E2E
	Visual-Semantic	HAD [197]	SC-MBM	Video-LLaMA	LLM4Brain [195]	FMA
	Mel-spectrograms	VGGSound [19]	VGGish	Audio Transformer	[128]	DPC
	Perceived Speech	Self-collected Dataset [203]	Transformer (Encoder)	-	neural2speech [203]	AG-EML
	Video	NSD [5]	CLIP	SDXL-Turbo	NeuroCreat [78]	FMA
EEG	Text	ZuCo [68]	Convolutional Transformer	Multi-View Transformer	EEG2TEXT [100]	E2E
			Multi-stream Transformer	-	CET-MAE [172]	FMA
		Things-EEG2 [57]	CLIP	GIT	PinVC [79]	FMA
	Silently Read Text	ZuCo [68]	VQ-VAE	BART	DeWave [35]	E2E
	Imagined Speech	ISD [182]	XI Model	XII Model	ETS [182]	E2E
	Imagined Speech	Self-collected Dataset [1]	LSTM-RNN	-	[1]	E2E
	Speech Recognition	Self-collected Dataset [58]	-	Whisper	[58]	E2E
	Speech Waveform	Self-collected Dataset [62]	-	DualGAN	Dual-DualGAN [62]	E2E
	Text Production	Self-collected Dataset [92]	CNN + Transformer	-	Brain2Qwerty [92]	DPC
	Speech Identification	MOUS [144]	wav2vec2	-	[29]	DPC
MEG	General MEG Features	UKB [115] HCP [36]	-	BERT-Adapted	BrainLM [18]	FMA
	Sentence Identification	MEG Narrative [8]	GAT	-	[56]	FMA
	Text Production	Self-collected Dataset [92]	CNN + Transformer	-	Brain2Qwerty [92]	DPC
	Speech Identification	MOUS [144]	wav2vec2	-	[29]	DPC
	Perceived Speech	Self-collected Dataset [203]	Transformer (Encoder)	-	neural2speech [203]	AG-EML
sEEG	Lexical Semantics	Self-collected Dataset [129]	FastText	-	[129]	FMA
	Speech Spectrograms	Self-collected Dataset [20]	Swin Transformer	-	SwinTW [20]	DPC
Intracranial EEG	Text	Self-collected Dataset [178]	RNN	-	[178]	E2E
ECoG	Speech Spectrograms	Self-collected Dataset [20]	Swin Transformer	-	SwinTW [20]	DPC
HD-DOT	Auditory Stimulus Info	HD-DOT[167]	Template Matching	-	[168]	FMA

encoding, and modality conversion, showing that an FM-style alignment layer can bridge disparate temporal/spatial scales and improve cross-subject generalization [45]. At the dataset level, recent large, paired M/EEG–fMRI resources and fusion analyses demonstrate consistent gains in modeling naturalistic visual processing when joint information is exploited, reinforcing the value of multi-neural supervision for both training and evaluation [112, 174]. In portable

settings, hybrid EEG+fNIRS architectures with cross-modal attention or graph fusion markedly outperform unimodal baselines on cognitive/affective tasks, suggesting a practical path toward out-of-lab BCIs that still benefit from FM-style invariances [97].

For neural–semantic alignment, multimodal decoders pair fused neural embeddings with strong generative priors. Visual decoders condition diffusion backbones on a joint latent that aggregates, for example, fMRI-derived spatial codes with MEG/EEG temporal embeddings, improving semantic retrieval and frame consistency for naturalistic stimuli; joint video–text embeddings further enhance encoding/decoding of movie stimuli versus uni-modal feature baselines [52]. Auxiliary behavioral streams can be folded in as well: MR-based eye tracking (DeepMReye) provides gaze priors directly from fMRI volumes, offering an inexpensive behavioral channel that stabilizes reconstructions and disambiguates scene context when combined with neural features [50]. Together, these designs operationalize a general recipe: map each stream (fMRI/MEG/EEG/fNIRS/behavior) into a shared FM-conditioned latent, perform attention or contrastive fusion with modality-dropout for robustness, and decode via conditional diffusion or VLM heads.

Looking ahead, multimodal decoding raises new opportunities and open problems: scalable co-training across sites and sensors; modality-adaptive schedulers that weight channels by SNR and task phase; and foundation models for neural data that natively support modality conversion (e.g., fMRI to MEG) and few-shot personalization. Early evidence from unified decoders and aligned representations indicates that fusing complementary signals consistently boosts semantic fidelity and temporal precision, delivering a coherent blueprint for FM-driven multimodal brain decoding in both lab-grade and portable settings.

5. Cross-Domain Trends and Breakthroughs

Across vision, language, and audio/speech decoding, pre-trained foundation models (FMs)—diffusion generators, CLIP-style dual encoders, large language models (LLMs), and self-supervised Transformers—have become pervasive as feature extractors, priors, and generators [59, 143, 128, 73]. A unifying blueprint is to first align brain activity with an FM latent space and then leverage the FM's generative or retrieval capacity for reconstruction or recognition. Alignment is achieved by (i) contrastive learning that places neural and stimulus embeddings in a shared manifold [93, 23], (ii) direct regression that predicts FM embeddings or conditioning variables from neural signals [123], or (iii) "encoding-model inversion," where a brain-prediction model guides an LLM-based search to select sequences whose predicted neural responses best match observed data [164].

Within this framework, FMs systematically compensate for modality-specific weaknesses. For fMRI, LLM priors impose linguistic structure over slow hemodynamics, enabling coherent semantic decoding, while keyframe-plus-flow strategies reduce temporal aliasing in video reconstruction [164, 59]. For EEG/MEG, strong visual priors from CLIP and diffusion models elevate semantic image reconstruction despite limited spatial precision and low SNR, and self-supervised learning extracts robust time-resolved features that generalize across tasks and subjects [93, 123, 73]. These advances coincide with a clear shift toward subject-agnostic and zero/few-shot regimes—e.g., MindLLM, SSL-trained MEG encoders, and zero-shot EEG visual decoding—aimed at reducing per-subject calibration and improving real-world viability [134, 73, 93, 164].

A second trajectory is deeper multimodality: systems increasingly align brain activity to vision-language models (VLMs), enabling flexible decoding into text and/or images within a single framework (e.g., UMBRAE) [134, 181]. In parallel, the community is probing temporal dynamics more explicitly in EEG/MEG, using FM-based analyses to characterize real-time processing, feature persistence, and semantic stabilization windows [12, 123, 93, 73]. Finally, generative pipelines themselves are becoming more sophisticated—multi-stage designs, multi-feature conditioning, and specialized architectures now dominate SOTA reconstructions—reflecting a maturation from simple stimulus-to-signal mappings to modular, prior-driven generation [59, 93, 123, 25].

6. Persisting Challenges and Future Horizons

Despite remarkable progress in leveraging foundation models (FMs) for brain-computer interfaces (BCIs), several major obstacles persist that limit real-world deployment, particularly in decoding speech and visual experiences from brain signals. This section integrates challenges and opportunities across modalities.

6.1. Current Hurdles for FMs in BCI

Despite rapid progress, foundation-model–based decoding faces several structural obstacles. First are computational and data constraints. State-of-the-art FMs demand substantial training and inference resources that are difficult to accommodate in wearable or clinical settings; latency and power budgets limit real-time use on embedded hardware. At the same time, effective pretraining and reliable evaluation require large, diverse, and well-annotated datasets, which are costly and slow to acquire in neuroscience. Speech tasks often lack sufficient labeled EEG/MEG, and visual decoding likewise depends on expensive high-resolution fMRI or specialized corpora (e.g., NSD, THINGS), constraining both model scaling and generalization [24, 76, 150].

A second hurdle arises from signal quality and modality trade-offs. EEG/MEG offer millisecond timing but suffer from low spatial specificity, non-stationarity, and vulnerability to artifacts; reliable access to deep sources is limited relative to fMRI. Conversely, fMRI localizes fine-grained spatial structure but blurs rapid dynamics due to

slow hemodynamics. Managing physiological and environmental noise—especially in EEG—remains non-trivial, and mapping complex stimuli such as natural language onto measured neural responses is rarely one-to-one, complicating model design and evaluation [151, 106].

Inter-subject variability further impedes robustness. Anatomical differences, functional topographies (e.g., retinotopy), state fluctuations, and session effects degrade cross-subject transfer; models trained on one individual often fail to generalize to others without costly calibration. Achieving zero-shot or strongly subject-agnostic performance via functional alignment or adaptation remains an open challenge across language and vision decoding alike [109, 24].

Model complexity and decoding ambiguity compound these difficulties. Open-vocabulary generation (free-form text or speech) is markedly harder than closed-set recognition, particularly for EEG. Higher-level semantic decoding is less mature than phonetic decoding, and imagined/silent speech (or internally generated imagery such as dreams) presents weaker, noisier signals without clear behavioral anchors. Architectures must also respect modality structure (e.g., irregular sensor graphs for EEG, 3D sampling for fMRI), while delivering low-latency, reliable outputs suitable for interactive BCIs. In parallel, debates persist about the cognitive plausibility of very large FMs as models of human processing and how best to constrain them for neuroscientific validity.

In addition, interpretability and evaluation remain limited. Many pipelines behave as black boxes, with unclear spatiotemporal mappings from brain activity to outputs; linear probes and attention maps provide partial insight but fall short of comprehensive explanations. Standard metrics (pixel-wise scores for images, BLEU/WER for language) can poorly reflect perceptual or semantic fidelity, underscoring the need for human-centered and task-relevant evaluations [59].

Finally, ethical considerations are central. Brain data are intrinsically sensitive, demanding stringent privacy, security, and governance to address "mind-reading" concerns and clarify ownership. FM training can amplify demographic and contextual biases, raising equity issues in performance and access. Robust frameworks for informed consent, user agency, and mental privacy boundaries are prerequisites for responsible deployment [150].

6.2. Opportunities and Future Research Directions

Meeting today's challenges points to a clear research agenda. A first priority is the development of Brain Foundation Models (BFMs): large, modality-aware encoders pretrained on diverse fMRI/EEG/MEG corpora to learn universal neural representations. Progress will hinge on self-supervised learning tailored to neural signals (masking, predictive objectives, cross-modal contrastive alignment), on hybrid architectures that bridge modalities (e.g., EEG to Text or EEG to Audio contrastive models), and on theoretically informed inductive biases such as predictive coding. Two complementary threads should run in parallel: systematically probing how scale, objective choice, and adaptation strategy affect encoding/decoding performance and cognitive plausibility; and leveraging powerful existing FMs (speech recognizers, LLMs, VLMs) as computational probes of brain representations.

A second frontier is decoding internally generated content. In speech, this includes silent speech and inner monologue; in vision, mental imagery and dreams. Because such phenomena lack overt behavioral anchors and unfold with weak, distributed signals, they will likely require new representational targets (semantic, affective, or episodic manifolds) and training curricula grounded in cognitive neuroscience—e.g., predictive coding or generative models that explicitly separate latent intent from sensory realization.

Third, advancing robustness and real-time BCIs remains essential. Methods must better tolerate noise, artifacts, and non-stationarity, while generalizing across subjects, tasks, datasets, and even modalities via functional alignment and adaptation. For EEG/MEG, rigorous comparisons between sensor-space and source-reconstructed decoding are needed; for open-vocabulary language and rich visual content, newer generative backbones may improve fidelity, controllability, and latency. The overarching goal is reliable, low-latency decoding suitable for interactive use in realistic settings.

Clinical translation is the fourth axis. Communication BCIs for individuals with ALS, stroke, or paralysis are a primary driver; success depends on robust models, streamlined calibration, and user-centered design (comfort, intuitive interfaces), alongside regulatory-grade reliability. Beyond assistive communication, decoding pipelines can serve basic neuroscience—clarifying language processing, semantic representation, and cognitive dynamics—and may yield biomarkers for neurological and psychiatric disorders. Emerging modalities such as HD-DOT add promising trade-offs for naturalistic language decoding and should be systematically evaluated.

Achieving these aims will also require efficient model design and innovative data strategies. Compression, distillation, parameter-efficient fine-tuning, and compact architectures can bring state-of-the-art decoding to edge and wearable hardware. On the data side, priorities include building large public datasets (e.g., via OpenNeuro), advancing augmentation for neurodata, expanding SSL on unlabeled recordings [73], and adopting privacy-preserving collection and training (federated/secure aggregation) to diversify cohorts without compromising confidentiality.

Generalization and personalization form a sixth pillar. Domain adaptation, meta-learning, and few-shot calibration can minimize per-subject effort; online adaptation can track drift; and cross-modal transfer can bootstrap under-resourced settings. Training on naturalistic data—continuous narratives, real-world scenes—should become standard to narrow the lab-to-life gap.

Interpretability and governance complete the road-map. XAI for BCI—from attention maps and feature attributions to causal perturbations and encoding-model probes—can

validate neuroscientific plausibility, surface spurious shortcuts, and build user trust. Bridging the lab-to-life gap further requires durable hardware, stable signal quality over months, and algorithms robust to long-term drift, all under ethical frameworks that foreground privacy, consent, equity, transparency, and accountability. Finally, treating FMs as tools for neuroscience—to instantiate and test hypotheses about neural information processing—offers a deep synergy: decoding improves as our models of the brain improve, and vice versa [72]. In the long term, reliable decoding of inner speech and abstract thought remains a moonshot goal, one that will likely emerge from the convergence of BFMs, theory-guided objectives, efficient adaptation, and ethically grounded, human-centered design [164, 4].

7. Conclusion

Foundation models have undeniably ushered in a new era for non-invasive brain decoding, fundamentally reshaping the field. By leveraging large-scale pre-training, transfer learning, and powerful generative capabilities, FMs demonstrate an unprecedented ability to interpret complex patterns in noisy signals from EEG, MEG, fMRI, and fNIRS. This survey highlighted breakthroughs: high-fidelity visual reconstruction (images/video) across modalities using CLIP and diffusion models; continuous semantic language decoding from fMRI via LLMs overcoming temporal limits, with progress towards subject-agnostic models; enhanced AAD performance using FM-inspired architectures; and emerging success in reconstructing complex sounds and music using generative approaches. A key theme is the synergistic use of FMs to actively compensate for inherent neuroimaging modality limitations, pushing the boundaries of non-invasive decoding.

However, the journey to widespread, reliable, and ethical deployment remains challenging. Computational/data requirements, interpretability needs, robust generalization (across subjects, tasks, environments), and profound ethical considerations persist. Continued innovation is needed not just in scaling models, but in developing efficient, adaptable, explainable, and personalized approaches tailored to BCI constraints.

The future lies at the intersection of AI, neuroscience, engineering, and ethics. Realizing FM's full potential requires parallel progress in sensing hardware, diverse datasets, realistic validation benchmarks, and proactive ethical dialogue. The ultimate goal is to harness FMs not only for powerful decoding tools but also to deepen our understanding of the brain, ensuring responsible development for human benefit. Continued interdisciplinary collaboration is paramount in transforming the promise of FM-driven non-invasive brain decoding into tangible improvements in health, communication, and human-world interaction.

Data availability

No data was used for the research described in the article. The dataset summarized in the article is shown in Table 2.

CRediT authorship contribution statement

Yifan Wang: Data curation, Investigation, Methodology, Visualization, Writing – original draft, Writing – review and editing. **Shaonan Wang:** Conceptualization, Supervision, Writing – original draft, Writing – review and editing. **Yunhao Zhang:** Methodology, Visualization, Writing – original draft. **Changde Du:** Writing – original draft. **Cunhang Fan:** Funding acquisition, Writing – original draft. **Dongyang Li:** Writing – original draft. **Hongpeng Zhou:** Writing – original draft. **Hongyu Zhang:** Writing – original draft. **Jixing Li:** Writing – original draft. **Quanying Liu:** Funding acquisition, Writing – original draft. **Wei Huang:** Funding acquisition, Writing – original draft. **Yizhuo Lu:** Writing – original draft. **Zijiao Chen:** Writing – original draft. **Jingyuan Sun:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review and editing.

Declaration of competing interest

Yifan Wang, Shaonan Wang, Changde Du, Cunhang Fan, Dongyang Li, Hongyu Zhang, Jixing Li, Quanying Liu, Wei Huang, Yizhuo Lu, Zijiao Chen and Jingyuan Sun declare that they have no conflict of interest or financial conflicts to disclose.

Acknowledgement

This work is supported by the STI 2030—Major Projects (No. 2021ZD0201500), the National Natural Science Foundation of China (NSFC) (No.62571002, 62406058), Excellent Youth Foundation of Anhui Scientific Committee (No. 2408085Y034); Sichuan Science and Technology Program (2025ZNSFSC1480); China Postdoctoral Science Foundation (2024M750360); Fundamental Research Funds for the Central Universities (ZYGX2024XJ051). Natural Science Foundation of China (62472206), National Key R&D Program of China (2025YFC3410000), Shenzhen Science and Technology Innovation Committee (RCYX202312 11090405003, KJZD20230923115221044). We acknowledge the support from the UMRI IDR Placement 2025 Pioneering project "Foundational AI Models for Brain Computer Interfaces and Neuro-Robotic Control", and The European High Performance Computing Joint Undertaking (EuroHPC JU) with project EHPC-DEV-2025D06-002 and project EHPC-BEN-2025B05-008.

References

[1] Abdulghani, M.M., Walters, W.L., Abed, K.H., 2023. Imagined speech classification using eeg and deep learning. Bioengineering 10, 649. https://www.mdpi.com/2306-5354/10/6/649.

[2] Abiri, R., Borhani, S., Sellers, E.W., Jiang, Y., Zhao, X., 2019. A comprehensive review of eeg-based brain–computer interface paradigms. Journal of neural engineering 16, 011001. https://iopscience.iop.org/article/10.1088/1741-2552/aaf12e/meta.

[3] Accou, B., Vanthornhout, J., hamme, H.V., Francart, T., 2023. Decoding of the speech envelope from eeg using the vlaai deep

neural network. Scientific Reports 13, 812. https://www.nature.com/articles/s41598-022-27332-2.

[4] Airhart, M., . Brain Activity Decoder Can Reveal Stories in People's Minds kernel description. https://cns.utexas.edu/news/podcast/brain-activity-decoder-can-reveal-stories-peoples-minds. Accessed: 2023-05-01.

[5] Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al., 2022. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience 25, 116–126. https://www.nature.com/articles/s41593-021-00962-x.

[6] Altaheri, H., Muhammad, G., Alsulaiman, M., Amin, S.U., Altuwaijri, G.A., Abdul, W., Bencherif, M.A., Faisal, M., 2023. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. Neural Computing and Applications 35, 14681–14722. https://link.springer.com/article/10.1007/s00521-021-06352-5.

[7] Antonello, R., Vaidya, A., Huth, A., 2023. Scaling laws for language encoding models in fmri. Advances in Neural Information Processing Systems 36, 21895–21907. https://proceedings.neurips.cc/paper_files/paper/2023/hash/4533e4a352440a32558c1c227602c323-Abstract-Conference.html.

[8] Armeni, K., Güçlü, U., van Gerven, M., Schoffelen, J.M., 2022. A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. Scientific Data 9, 278. https://www.nature.com/articles/s41597-022-01382-7.

[9] Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems 33, 12449–12460.

[10] Bai, Y., Wang, X., Cao, Y.p., Ge, Y., Yuan, C., Shan, Y., 2023. Dreamdiffusion: Generating high-quality images from brain eeg signals. arXiv preprint arXiv:2306.16934 https://arxiv.org/abs/2306.16934.

[11] Beliy, R., Gaziv, G., Hoogi, A., Strappini, F., Golan, T., Irani, M., 2019. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. Advances in Neural Information Processing Systems 32. https://proceedings.neurips.cc/paper_files/paper/2019/hash/7d2be41b1bde6ff8fe45150c37488ebb-Abstract.html.

[12] Benchetrit, Y., Banville, H., King, J.R., 2023. Brain decoding: toward real-time reconstruction of visual perception. arXiv preprint arXiv:2310.19812 https://arxiv.org/abs/2310.19812.

[13] Buzsáki, G., Anastassiou, C.A., Koch, C., 2012. The origin of extracellular fields and currents—eeg, ecog, lfp and spikes. Nature reviews neuroscience 13, 407–420. https://www.nature.com/articles/nrn3241.

[14] Cai, S., Schultz, T., Li, H., 2023. Brain topology modeling with eeg-graphs for auditory spatial attention detection. IEEE Transactions on Biomedical Engineering 71, 171–182. https://ieeexplore.ieee.org/abstract/document/10178086/.

[15] Cai, S., Su, E., Song, Y., Xie, L., Li, H., 2020. Low latency auditory attention detection with common spatial pattern analysis of eeg signals., in: INTERSPEECH, pp. 2772–2776. http://www.interspeech2020.org/uploadfile/pdf/Wed-SS-2-7-3.pdf.

[16] Cao, X., Gong, P., Zhang, L., Zhang, D., 2025. Eeg-clip: A transformer-based framework for eeg-guided image generation. Neural Networks , 108167.

[17] Carìa, A., 2025. Integrating large language models and brain decoding for augmented human-computer interaction: A prototype llm-p3-bci speller, in: Future of Information and Communication Conference, Springer. pp. 403–416. https://link.springer.com/chapter/10.1007/978-3-031-84457-7_25.

[18] Caro, J.O., Fonseca, A.H.d.O., Averill, C., Rizvi, S.A., Rosati, M., Cross, J.L., Mittal, P., Zappala, E., Levine, D., Dhodapkar, R.M., et al., 2023. Brainlm: A foundation model for brain activity recordings. bioRxiv , 2023–09https://www.biorxiv.org/content/10.1101/2023.09.12.557460v2.abstract.

[19] Chen, H., Xie, W., Vedaldi, A., Zisserman, A., 2020. Vggsound: A large-scale audio-visual dataset, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 721–725. https://ieeexplore.ieee.org/abstract/document/9053174.

[20] Chen, J., Chen, X., Wang, R., Le, C., Khalilian-Gourtani, A., Jensen, E., Dugan, P., Doyle, W., Devinsky, O., Friedman, D., et al., 2024a. Subject-agnostic transformer-based neural speech decoding from surface and depth electrode signals. bioRxiv https://pmc.ncbi.nlm.nih.gov/articles/PMC10980022/.

[21] Chen, J., Qi, Y., Wang, Y., Pan, G., 2025a. Bridging the gap between brain and machine in interpreting visual semantics: Towards self-adaptive brain-to-text decoding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21938–21948.

[22] Chen, J., Qi, Y., Wang, Y., Pan, G., 2025b. Mindgpt: Interpreting what you see with non-invasive brain recordings. IEEE Transactions on Image Processing .

[23] Chen, X., Du, C., Liu, C., Wang, Y., He, H., 2024b. Open-vocabulary auditory neural decoding using fmri-prompted llm. arXiv preprint arXiv:2405.07840 https://arxiv.org/abs/2405.07840.

[24] Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H., 2023a. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding , 22710–22720https://arxiv.org/abs/2211.06956.

[25] Chen, Z., Qing, J., Zhou, J.H., 2023b. Cinematic mindscapes: High-quality video reconstruction from brain activity. Advances in Neural Information Processing Systems 36, 24841–24858. https://proceedings.neurips.cc/paper_files/paper/2023/hash/4e5e0daf4b05d8bfc6377f33fd53a8f4-Abstract-Conference.html.

[26] Craik, A., He, Y., Contreras-Vidal, J.L., 2019. Deep learning for electroencephalogram (eeg) classification tasks: a review. Journal of neural engineering 16, 031001. https://iopscience.iop.org/article/10.1088/1741-2552/ab0ab5/meta.

[27] De Cheveigné, A., Wong, D.D., Di Liberto, G.M., Hjortkjær, J., Slaney, M., Lalor, E., 2018. Decoding the auditory brain with canonical component analysis. NeuroImage 172, 206–216. https://www.sciencedirect.com/science/article/pii/S1053811918300338.

[28] De Taillez, T., Kollmeier, B., Meyer, B.T., 2020. Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. European Journal of Neuroscience 51, 1234–1241. https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.13790.

[29] Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., King, J.R., 2023. Decoding speech perception from non-invasive brain recordings. Nature Machine Intelligence 5, 1097–1107. https://www.nature.com/articles/s42256-023-00714-5.

[30] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255. https://ieeexplore.ieee.org/document/5206848.

[31] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding , 4171–4186https://arxiv.org/abs/1810.04805.

[32] Dmochowski, J.P., Ki, J.J., DeGuzman, P., Sajda, P., Parra, L.C., 2018. Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity. NeuroImage 180, 134–146. https://www.sciencedirect.com/science/article/pii/S1053811917304299.

[33] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

[34] Du, B., Cheng, X., Duan, Y., Ning, H., 2022. fmri brain decoding and its applications in brain–computer interface: A survey. Brain Sciences 12, 228.

[35] Duan, Y., Zhou, J., Wang, Z., Wang, Y.K., Lin, C.T., 2023. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. arXiv preprint arXiv:2309.14030 https://arxiv.org/abs/2309.

14030.

[36] Elam, J.S., Glasser, M.F., Harms, M.P., Sotiropoulos, S.N., Andersson, J.L., Burgess, G.C., Curtiss, S.W., Oostenveld, R., Larson-Prior, L.J., Schoffelen, J.M., et al., 2021. The human connectome project: a retrospective. NeuroImage 244, 118543. https://www.sciencedirect.com/science/article/pii/S1053811921008168.

[37] Fan, C., Yang, X., Zhang, H., Chen, Y., Li, L., Zhou, J., Lv, Z., 2025a. Listennet: A lightweight spatio-temporal enhancement nested network for auditory attention detection, in: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25, pp. 4137–4145. https://doi.org/10.24963/ijcai.2025/461.

[38] Fan, C., Zhang, H., Huang, W., Xue, J., Tao, J., Yi, J., Lv, Z., Wu, X., 2024. Dgsd: Dynamical graph self-distillation for eeg-based auditory spatial attention detection. Neural Networks 179, 106580. https://www.sciencedirect.com/science/article/abs/pii/S0893608024005045.

[39] Fan, C., Zhang, H., Ni, Q., Zhang, J., Tao, J., Zhou, J., Yi, J., Lv, Z., Wu, X., 2025b. Seeing helps hearing: A multi-modal dataset and a mamba-based dual branch parallel network for auditory attention decoding. Information Fusion , 102946https://www.sciencedirect.com/science/article/abs/pii/S1566253525000193.

[40] Fan, C., Zhang, S., Zhang, J., Liu, E., Li, X., Zhao, G., Lv, Z., 2025c. Dmf2mel: A dynamic multiscale fusion network for eeg-driven mel spectrogram reconstruction, in: Proceedings of the 33rd ACM International Conference on Multimedia, pp. 6977–6985. https://doi.org/10.1145/3746027.3755501.

[41] Fei, T., Uppal, A., Jackson, I.L., Wang, D., de Sa, V.R., . Perceptogram: Visual reconstruction from eeg using image generative models https://openreview.net/forum?id=IZOeRDS6zU.

[42] Feng, F., Nie, J., 2025. From eye to mind: brain2text decoding reveals the neural mechanisms of visual semantic processing. arXiv preprint arXiv:2503.22697 https://arxiv.org/abs/2503.22697.

[43] Ferrante, M., Boccato, T., Bargione, S., Toschi, N., 2024a. Decoding visual brain representations from electroencephalography through knowledge distillation and latent diffusion models. Computers in Biology and Medicine 178, 108701. https://www.sciencedirect.com/science/article/pii/S0010482524007868.

[44] Ferrante, M., Boccato, T., Passamonti, L., Toschi, N., 2024b. Retrieving and reconstructing conceptually similar images from fmri with latent diffusion models and a neuro-inspired brain decoding model. Journal of Neural Engineering 21, 046001. https://iopscience.iop.org/article/10.1088/1741-2552/ad593c/meta.

[45] Ferrante, M., Boccato, T., Rashkov, G., Toschi, N., 2024c. Towards neural foundation models for vision: Aligning eeg, meg, and fmri representations for decoding, encoding, and modality conversion. arXiv preprint arXiv:2411.09723 https://arxiv.org/abs/2411.09723.

[46] Ferrante, M., Boccato, T., Toschi, N., 2022. Semantic brain decoding: from fmri to conceptually similar image reconstruction of visual stimuli. arXiv preprint arXiv:2212.06726 https://arxiv.org/abs/2212.06726.

[47] Ferrante, M., Ozcelik, F., Boccato, T., VanRullen, R., Toschi, N., 2023. Brain captioning: Decoding human brain activity into images and text. arXiv preprint arXiv:2305.11560 https://arxiv.org/abs/2305.11560.

[48] Ferrari, M., Quaresima, V., 2012. A brief review on the history of human functional near-infrared spectroscopy (fnirs) development and fields of application. Neuroimage 63, 921–935. https://www.sciencedirect.com/science/article/abs/pii/S1053811912003308.

[49] Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., Obleser, J., 2017. Single-channel in-ear-eeg detects the focus of auditory attention to concurrent tone streams and mixed speech. Journal of neural engineering 14, 036020. https://iopscience.iop.org/article/10.1088/1741-2552/aa66dd/meta.

[50] Frey, M., Nau, M., Doeller, C.F., 2021. Magnetic resonance-based eye tracking using deep neural networks. Nature neuroscience 24, 1772–1779.

[51] Fu, H., Wang, H., Chin, J.J., Shen, Z., 2025a. Brainvis: Exploring the bridge between brain and visual signals via image reconstruction, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5.

[52] Fu, M., Chen, G., Zhang, Y., Zhang, M., Wang, Y., 2025b. Comprehensive neural representations of naturalistic stimuli through multimodal deep learning. bioRxiv , 2025–04.

[53] Gao, C., Ma, Z., Chen, J., Li, P., Huang, S., Li, J., 2025a. Scaling, but not instruction tuning, increases large language models' alignment with language processing in the human brain. Nature Computational Science https://www.biorxiv.org/content/10.1101/2024.08.15.608196v4.abstract.

[54] Gao, J., Fu, Y., Fu, Y., Wang, Y., Qian, X., Feng, J., 2025b. Mind-3d++: Advancing fmri-based 3d reconstruction with high-quality textured mesh generation and a comprehensive dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence .

[55] Geirnaert, S., Vandecappelle, S., Alickovic, E., De Cheveigne, A., Lalor, E., Meyer, B.T., Miran, S., Francart, T., Bertrand, A., 2021. Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices. IEEE Signal Processing Magazine 38, 89–102. https://ieeexplore.ieee.org/abstract/document/9467380.

[56] Gideoni, Y., Timms, R.C., Jones, O.P., 2024. Non-invasive neural decoding in source reconstructed brain space. arXiv preprint arXiv:2410.19838 https://arxiv.org/abs/2410.19838.

[57] Gifford, A.T., Dwivedi, K., Roig, G., Cichy, R.M., 2022. A large and rich eeg dataset for modeling human visual object recognition. NeuroImage 264, 119754. https://www.sciencedirect.com/science/article/pii/S1053811922008758.

[58] Goldstein, A., Wang, H., Niekerken, L., Zada, Z., Aubrey, B., Sheffer, T., Nastase, S.A., Gazula, H., Schain, M., Singh, A., et al., 2023. Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations. bioRxiv , 2023–06https://www.biorxiv.org/content/10.1101/2023.06.26.546557v1.abstract.

[59] Gong, Z., Bao, G., Zhang, Q., Wan, Z., Miao, D., Wang, S., Zhu, L., Wang, C., Xu, R., Hu, L., et al., 2024. Neuroclips: Towards high-fidelity and smooth fmri-to-video reconstruction. Advances in Neural Information Processing Systems 37, 51655–51683. https://proceedings.neurips.cc/paper_files/paper/2024/hash/5c594bf6223b67109441c9e0c97542ed-Abstract-Conference.html.

[60] Gong, Z., Zhang, Q., Bao, G., Zhu, L., Xu, R., Liu, K., Hu, L., Miao, D., 2025. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 14247–14255.

[61] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Communications of the ACM 63, 139–144.

[62] Guo, Y., Liu, T., Zhang, X., Wang, A., Wang, W., 2023. End-to-end translation of human neural activity to speech with a dual–dual generative adversarial network. Knowledge-based systems 277, 110837. https://www.sciencedirect.com/science/article/pii/S0950705123005877.

[63] Han, K., Wen, H., Shi, J., Lu, K.H., Zhang, Y., Fu, D., Liu, Z., 2019. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. NeuroImage 198, 125–136. https://www.sciencedirect.com/science/article/abs/pii/S1053811919304318.

[64] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners , 16000–16009https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.

[65] Hebart, M.N., Contier, O., Teichmann, L., Rockter, A.H., Zheng, C.Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., Baker, C.I., 2023. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. Elife 12, e82580. https://elifesciences.org/articles/82580.

[66] Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T.B., Dhariwal, P., Gray, S., et al., 2020. Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701 https://arxiv.org/abs/2010.14701.

[67] Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A., Yang, Y., Zhou, Y., 2017. Deep learning scaling is predictable. Empirically. arXiv 1712, 2. https://arxiv.org/abs/1712.00409.

[68] Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., Langer, N., 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. Scientific data 5, 1–13. https://www.nature.com/articles/sdata2018291.

[69] Horikawa, T., Kamitani, Y., 2017. Generic decoding of seen and imagined objects using hierarchical visual features. Nature communications 8, 15037. https://www.nature.com/articles/ncomms15037.

[70] Huang, W., 2024. Brainchat: Decoding semantic information from fmri using vision-language pretrained models. arXiv preprint arXiv:2406.07584 https://arxiv.org/abs/2406.07584.

[71] Huang, W., Ma, K., Xie, T., Wang, H., 2025. Brainchat: Interactive semantic information decoding from fmri using large-scale vision-language pretrained models, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5.

[72] Huth, A.G., De Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. https://www.nature.com/articles/nature17637.

[73] Jayalath, D., Landau, G., Shillingford, B., Woolrich, M., Jones, O.P., 2024. The brain's bitter lesson: Scaling speech decoding with self-supervised learning. arXiv preprint arXiv:2406.04328 https://arxiv.org/abs/2406.04328.

[74] Jayaram, V., Barachant, A., 2018. Moabb: trustworthy algorithm benchmarking for bcis. Journal of neural engineering 15, 066011.

[75] Jiang, S., Li, H., Meng, Z., Su, F., Zhao, Z., 2025. Mind2word: Towards generalized visual neural representations for high-quality video reconstruction. Expert Systems with Applications , 129154.

[76] Jiang, W.B., Zhao, L.M., Lu, B.L., 2024. Large brain model for learning generic representations with tremendous eeg data in bci. arXiv preprint arXiv:2405.18765 https://arxiv.org/abs/2405.18765.

[77] Jiao, Z., You, H., Yang, F., Li, X., Zhang, H., Shen, D., 2019. Decoding eeg by visual-guided deep neural networks., in: IJCAI, Macao. pp. 1387–1393. https://www.ijcai.org/proceedings/2019/0192.pdf.

[78] Jing, H., Jiang, D., Ma, Y., Hua, H., Huang, B., Zheng, N., 2025a. Beyond brain decoding: Visual-semantic reconstructions to mental creation extension based on fmri, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 19258–19268.

[79] Jing, H., Ma, Y., Yang, P., Hua, H., Zheng, N., 2025b. Pinpointing visual content: Disentangled features in multimodal model for eeg representation learning and decoding. Knowledge-Based Systems , 114212.

[80] Kalantari, F., Faez, K., Amindavar, H., Nazari, S., 2025. Improved image reconstruction from brain activity through automatic image captioning. Scientific Reports 15, 4907. https://www.nature.com/articles/s41598-025-89242-3.

[81] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D., 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 https://arxiv.org/pdf/2001.08361/1000.

[82] Kavasidis, I., Palazzo, S., Spampinato, C., Giordano, D., Shah, M., 2017. Brain2image: Converting brain signals into images, in: Proceedings of the 25th ACM international conference on Multimedia, pp. 1809–1817. https://dl.acm.org/doi/abs/10.1145/3123266.3127907.

[83] Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452, 352–355. https://www.nature.com/articles/nature06713.

[84] Khare, S., Choubey, R.N., Amar, L., Udutalapalli, V., 2022. Neurovision: perceived image regeneration using cprogan. Neural Computing and Applications 34, 5979–5991. https://link.springer.com/article/10.1007/s00521-021-06774-1.

[85] Kim, B.H., Choi, J., Yun, E., Kim, K., Li, X., Lee, J., 2023. Large-scale graph representation learning of dynamic brain connectome with transformers. arXiv preprint arXiv:2312.14939 .

[86] Kim, J.M., Choi, W., Park, S.J., Heo, K.S., Son, Y.H., Oh, J.H., Shin, D.H., Kam, T.E., 2025. Seeeeg: Semantic-aware eeg-based multi-modal retrieval-augmented generation for high-fidelity visual brain decoding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4824–4833.

[87] Kupershmidt, G., Beliy, R., Gaziv, G., Irani, M., 2022. A penny for your (visual) thoughts: Self-supervised reconstruction of natural movies from brain activity. arXiv preprint arXiv:2206.03544 https://arxiv.org/abs/2206.03544.

[88] Lan, Y.T., Ren, K., Wang, Y., Zheng, W.L., Li, D., Lu, B.L., Qiu, L., 2023. Seeing through the brain: image reconstruction of visual perception from human brain signals. arXiv preprint arXiv:2308.02510 https://arxiv.org/abs/2308.02510.

[89] Lang, K., Fang, J., Su, G., 2025. Neurofusionnet: cross-modal modeling from brain activity to visual understanding. Frontiers in Computational Neuroscience 19, 1545971.

[90] Le, L., Ambrogioni, L., Seeliger, K., Güçlütürk, Y., van Gerven, M., Güçlü, U., 2022. Brain2pix: Fully convolutional naturalistic video frame reconstruction from brain activity. Frontiers in Neuroscience 16, 940972. https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.940972/full.

[91] LeBel, A., Wagner, L., Jain, S., Adhikari-Desai, A., Gupta, B., Morgenthal, A., Tang, J., Xu, L., Huth, A.G., 2023. A natural language fmri dataset for voxelwise encoding models. Scientific Data 10, 555. https://www.nature.com/articles/s41597-023-02437-z.

[92] Lévy, J., Zhang, M., Pinet, S., Rapin, J., Banville, H., d'Ascoli, S., King, J.R., 2025. Brain-to-text decoding: A non-invasive approach via typing. arXiv preprint arXiv:2502.17480 https://arxiv.org/abs/2502.17480.

[93] Li, D., Wei, C., Li, S., Zou, J., Qin, H., Liu, Q., 2024. Visual decoding and reconstruction via eeg embeddings with guided diffusion. arXiv preprint arXiv:2403.07721 https://arxiv.org/abs/2403.07721.

[94] Li, H., Wu, H., Chen, B., 2025a. Neuraldiffuser: Neuroscience-inspired diffusion guidance for fmri visual reconstruction. IEEE Transactions on Image Processing .

[95] Li, J., Li, D., Xiong, C., Hoi, S., 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, PMLR. pp. 12888–12900.

[96] Li, L., Fan, C., Zhang, H., Zhang, J., Yang, X., Zhou, J., Lv, Z., 2025b. Mhanet: Multi-scale hybrid attention network for auditory attention detection, in: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25, pp. 4173–4181. https://doi.org/10.24963/ijcai.2025/465.

[97] Li, Y., Zhu, L., Huang, A., Zhang, J., Yuan, P., 2025c. Multimodal mbc-att: cross-modality attentional fusion of eeg-fnirs for cognitive state decoding. Frontiers in Human Neuroscience 19, 1660532.

[98] Lin, S., Sprague, T., Singh, A.K., 2022. Mind reader: Reconstructing complex images from brain activities. Advances in Neural Information Processing Systems 35, 29624–29636. https://arxiv.org/abs/2210.01769.

[99] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13, Springer. pp. 740–755. https://arxiv.org/abs/1405.0312.

[100] Liu, H., Hajialigol, D., Antony, B., Han, A., Wang, X., 2024a. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pre-training and multi-view transformer. arXiv preprint arXiv:2405.02165 https://arxiv.org/abs/2405.02165.

[101] Liu, P., Dong, G., Guo, D., Li, K., Li, F., Yang, X., Wang, M., Ying, X., 2025a. A survey on fmri-based brain decoding for reconstructing multimodal stimuli. arXiv preprint arXiv:2503.15978 https://arxiv.org/abs/2503.15978.

[102] Liu, X.H., Liu, Y.K., Wang, Y., Ren, K., Shi, H., Wang, Z., Li, D., Lu, B.L., Zheng, W.L., 2024b. Eeg2video: Towards decoding dynamic visual perception from eeg signals. Advances in Neural Information Processing Systems 37, 72245–72273. https://proceedings.neurips.cc/paper_files/paper/2024/hash/84bad835faaf48f24d990072bb5b80ee-Abstract-Conference.html.

[103] Liu, X.H., Lu, B.L., Zheng, W.L., 2025b. Eegmirror: Leveraging eeg data in the wild via montage-agnostic self-supervision for eeg to video decoding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 18273–18283.

[104] Liu, Y., Ma, Y., Zhou, W., Zhu, G., Zheng, N., 2023. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding. arXiv preprint arXiv:2302.12971 https://arxiv.org/abs/2302.12971.

[105] Liu, Y., Ma, Y., Zhu, G., Jing, H., Zheng, N., 2025c. See through their minds: Learning transferable brain decoding models from cross-subject fmri, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5730–5738.

[106] Logothetis, N.K., 2008. What we can do and what we cannot do with fmri. Nature 453, 869–878. https://www.nature.com/articles/nature06976.

[107] Lopez, E., Sigillo, L., Colonnese, F., Panella, M., Comminiello, D., 2025. Guess what i think: Streamlined eeg-to-image generation with latent diffusion models, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5.

[108] Lu, Y., Du, C., Zhou, Q., Wang, D., He, H., 2023. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion, in: Proceedings of the 31st ACM International Conference on Multimedia, pp. 5899–5908. https://dl.acm.org/doi/abs/10.1145/3581783.3613832.

[109] Luo, Y., Liu, W., Li, H., Lu, Y., Lu, B.L., 2024. A cross-scenario and cross-subject domain adaptation method for driving fatigue detection. Journal of Neural Engineering 21, 046004. https://iopscience.iop.org/article/10.1088/1741-2552/ad546d/meta.

[110] Mai, W., Zhang, Z., 2023. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. arXiv preprint arXiv:2308.07428 https://arxiv.org/abs/2308.07428.

[111] McGeady, C., Vučković, A., Puthusserypady, S., 2019. A hybrid mi-ssvep based brain computer interface for potential upper limb neurorehabilitation: a pilot study, in: 2019 7th International Winter Conference on Brain-Computer Interface (BCI), IEEE. pp. 1–6. https://ieeexplore.ieee.org/abstract/document/8737333.

[112] McNabb, C.B., Driver, I.D., Hyde, V., Hughes, G., Chandler, H.L., Thomas, H., Allen, C., Messaritaki, E., Hodgetts, C.J., Hedge, C., et al., 2025. Wand: A multi-modal dataset integrating advanced mri, meg, and tms for multi-scale brain analysis. Scientific Data 12, 220.

[113] Mehmood, T., Ahmad, H., Shakeel, M.H., Taj, M., 2025. Catvis: Context-aware thought visualization, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 98–108.

[114] Meng, L., Yang, C., 2023. Dual-guided brain diffusion model: Natural image reconstruction from human visual stimulus fmri. Bioengineering 10, 1117. https://www.nature.com/articles/s41598-023-42891-8.

[115] Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., et al., 2016. Multimodal population brain imaging in the uk biobank prospective epidemiological study. Nature neuroscience 19, 1523–1536. https://www.nature.com/articles/nn.4393.

[116] Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., Just, M.A., 2008. Predicting human brain activity associated with the meanings of nouns. science 320, 1191–1195. https://www.science.org/doi/abs/10.1126/science.1152876.

[117] Mozafari, M., Reddy, L., VanRullen, R., 2020. Reconstructing natural scenes from fmri patterns using bigbigan, in: 2020 International joint conference on neural networks (IJCNN), IEEE. pp. 1–8. https://ieeexplore.ieee.org/abstract/document/9206960.

[118] Nakai, T., Koide-Majima, N., Nishimoto, S., 2021. Correspondence of categorical and feature-based representations of music in the human brain. Brain and behavior 11, e01936. https://onlinelibrary.wiley.com/doi/full/10.1002/brb3.1936.

[119] Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., Gallant, J.L., 2009. Bayesian reconstruction of natural images from human brain activity. Neuron 63, 902–915. https://www.cell.com/neuron/fulltext/S0896-6273(09)00685-0.

[120] Nastase, S.A., Liu, Y.F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C.J., Yeshurun, Y., Regev, M., et al., 2021. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. Scientific data 8, 250. https://www.nature.com/articles/s41597-021-01033-3.

[121] Ni, Q., Zhang, H., Fan, C., Pei, S., Zhou, C., Lv, Z., 2024. Dbpnet: dual-branch parallel network with temporal-frequency fusion for auditory attention detection, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, pp. 3115–3123. https://doi.org/10.24963/ijcai.2024/345.

[122] Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Current biology 21, 1641–1646. https://www.sciencedirect.com/science/article/pii/S0960982211009377.

[123] NoAILabs, . Brain decoding, reconstruction of visual perception. Accessed: 2024-02-21.

[124] Ozcelik, F., Choksi, B., Mozafari, M., Reddy, L., VanRullen, R., 2022. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans , 1–8https://ieeexplore.ieee.org/abstract/document/9892673.

[125] Ozcelik, F., VanRullen, R., 2023. Natural scene reconstruction from fmri signals using generative latent diffusion. Scientific Reports 13, 15666. https://www.nature.com/articles/s41598-023-42891-8.

[126] Pais-Vieira, M., Lebedev, M., Kunicki, C., Wang, J., Nicolelis, M.A., 2013. A brain-to-brain interface for real-time sharing of sensorimotor information. Scientific reports 3, 1319. https://www.nature.com/articles/srep01319.

[127] Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D., Shah, M., 2017. Generative adversarial networks conditioned by brain signals, in: Proceedings of the IEEE international conference on computer vision, pp. 3410–3418. https://openaccess.thecvf.com/content_iccv_2017/html/Palazzo_Generative_Adversarial_Networks_ICCV_2017_paper.html.

[128] Park, J.Y., Tsukamoto, M., Tanaka, M., Kamitani, Y., 2023. Sound reconstruction from human brain activity via a generative model with brain-like auditory features. arXiv preprint arXiv:2306.11629 https://arxiv.org/abs/2306.11629.

[129] Pescatore, C.R., Zhang, H., Hadjinicolaou, A.E., Paulk, A.C., Rolston, J.D., Richardson, R.M., Williams, Z.M., Cai, J., Cash, S.S., 2025. Decoding semantics from natural speech using human intracranial eeg. bioRxiv , 2025–02https://www.biorxiv.org/content/10.1101/2025.02.10.637051v1.abstract.

[130] Postolache, E., Polouliakh, N., Kitano, H., Connelly, A., Rodolà, E., Cosmo, L., Akama, T., 2025. Naturalistic music decoding from eeg data via latent diffusion models, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5. https://ieeexplore.ieee.org/abstract/document/10887735.

[131] Pritchard, W.S., 1981. Psychophysiology of p300. Psychological bulletin 89, 506. https://psycnet.apa.org/record/1981-20721-001.

[132] Puffay, C., Accou, B., Bollens, L., Monesi, M.J., Vanthornhout, J., Francart, T., et al., 2023. Relating eeg to continuous speech using deep neural networks: a review. Journal of Neural Engineering 20, 041003. https://iopscience.iop.org/article/10.1088/1741-2552/

ace73f/meta.

[133] Qian, X., Wang, Y., Fu, Y., Sun, X., Xue, X., Feng, J., 2023. Joint fmri decoding and encoding with latent embedding alignment. arXiv preprint arXiv:2303.14730 https://arxiv.org/abs/2303.14730.

[134] Qiu, W., Huang, Z., Hu, H., Feng, A., Yan, Y., Ying, R., 2025a. Mindllm: A subject-agnostic and versatile model for fmri-to-text decoding. arXiv preprint arXiv:2502.15786 https://arxiv.org/abs/2502.15786.

[135] Qiu, Y., Liu, H., Zhao, M., 2025b. A review of brain–computer interface-based language decoding: From signal interpretation to intelligent communication. Applied Sciences 15, 392. https://www.mdpi.com/2076-3417/15/1/392.

[136] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision , 8748–8763https://proceedings.mlr.press/v139/radford21a.

[137] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9. https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf.

[138] Rakhimberdina, Z., Jodelet, Q., Liu, X., Murata, T., 2021. Natural image reconstruction from fmri using deep learning: A survey. Frontiers in neuroscience 15, 795488.

[139] Ren, Z., Li, J., Xue, X., Li, X., Yang, F., Jiao, Z., Gao, X., 2021. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. NeuroImage 228, 117602. https://www.sciencedirect.com/science/article/pii/S1053811920310879.

[140] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695.

[141] Rossi, S., Antal, A., Bestmann, S., Bikson, M., Brewer, C., Brockmöller, J., Carpenter, L.L., Cincotta, M., Chen, R., Daskalakis, J.D., et al., 2021. Safety and recommendations for tms use in healthy subjects and patient populations, with updates on training, ethical and regulatory issues: Expert guidelines. Clinical Neurophysiology 132, 269–306. https://www.sciencedirect.com/science/article/pii/S1388245720305149.

[142] Rubin, T.N., Koyejo, O., Gorgolewski, K.J., Jones, M.N., Poldrack, R.A., Yarkoni, T., 2017. Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition. PLoS computational biology 13, e1005649. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005649.

[143] Sato, A., Kobayashi, I., 2025. Decoding semantic representations in the brain under language stimuli with large language models, in: Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025), pp. 53–67. https://aclanthology.org/2025.wraicogs-1.6/.

[144] Schoffelen, J.M., Oostenveld, R., Lam, N.H., Uddén, J., Hultén, A., Hagoort, P., 2019. A 204-subject multimodal neuroimaging dataset to study language processing. Scientific data 6, 17. https://www.nature.com/articles/s41597-019-0020-y.

[145] Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., Norman, K., et al., 2024. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. Advances in Neural Information Processing Systems 36. https://medarc-ai.github.io/mindeye/.

[146] Seeliger, K., Sommers, R., Güçlü, U., Bosch, S., Van Gerven, M., 2019. A large single-participant fmri dataset for probing brain responses to naturalistic stimuli in space and time. BioRxiv , 687681https://www.biorxiv.org/content/10.1101/687681v1.

[147] Shehzad, A., Yu, S., Zhang, D., Abid, S., Cheng, X., Zhou, J., Xia, F., 2024. Braingt: Multifunctional brain graph transformer for brain disorder diagnosis. medRxiv , 2024–08.

[148] Shen, G., Dwivedi, K., Majima, K., Horikawa, T., Kamitani, Y., 2019. End-to-end deep image reconstruction from human brain activity. Frontiers in computational neuroscience 13, 21. https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2019.00021/full.

[149] Shen, V., Kunanbayev, K., Kim, D.S., 2025. Decoding fmri data into captions using prefix language modeling. arXiv preprint arXiv:2501.02570 https://arxiv.org/abs/2501.02570.

[150] Shirakawa, K., Nagano, Y., Tanaka, M., Aoki, S.C., Majima, K., Muraki, Y., Kamitani, Y., 2024. Spurious reconstruction from brain activity. arXiv preprint arXiv:2405.10078 https://arxiv.org/abs/2405.10078.

[151] da Silva, F.L., 2013. Eeg and meg: relevance to neuroscience. Neuron 80, 1112–1128. https://www.cell.com/neuron/fulltext/S0896-6273(13)00920-3?script=true&mobileUi=0.

[152] Simon, A.M.D., Østergaard, J., Bech, S., Loquet, G.S.J.M., 2022. Optimal time lags for linear cortical auditory attention detection: differences between speech and music listening, in: 19th International Symposium on Hearing: Psychoacoustics, Physiology of Hearing, and Auditory Modelling, from the Ear to the Brain, Zenodo. https://vbn.aau.dk/en/publications/optimal-time-lags-for-linear-cortical-auditory-attention-detectio.

[153] Solon, A.J., Lawhern, V.J., Touryan, J., McDaniel, J.R., Ries, A.J., Gordon, S.M., 2019. Decoding p300 variability using convolutional neural networks. Frontiers in human neuroscience 13, 201. https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2019.00201/full.

[154] Song, Y., Liu, B., Li, X., Shi, N., Wang, Y., Gao, X., 2023. Decoding natural images from eeg for object recognition. arXiv preprint arXiv:2308.13234 https://arxiv.org/abs/2308.13234.

[155] Spampinato, C., Palazzo, S., Kavasidis, I., Giordano, D., Souly, N., Shah, M., 2017. Deep learning human mind for automated visual classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6809–6817. http://openaccess.thecvf.com/content_cvpr_2017/html/Spampinato_Deep_Learning_Human_CVPR_2017_paper.html.

[156] Su, E., Cai, S., Li, P., Xie, L., Li, H., 2021. Auditory attention detection with eeg channel attention, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE. pp. 5804–5807. https://ieeexplore.ieee.org/abstract/document/9630508.

[157] Sun, J., Li, M., Chen, Z., Moens, M.F., 2024. Neurocine: Decoding vivid video sequences from human brain activties. arXiv preprint arXiv:2402.01590 .

[158] Sun, J., Li, M., Chen, Z., Zhang, Y., Wang, S., Moens, M.F., 2023a. Contrast, attend and diffuse to decode high-resolution images from brain activities. Advances in Neural Information Processing Systems 36, 12332–12348. https://proceedings.neurips.cc/paper_files/paper/2023/hash/28dad4a70f748a2980998d3ed0f1b8d2-Abstract-Conference.html.

[159] Sun, J., Li, M., Moens, M.F., 2023b. Decoding realistic images from brain activity with contrastive self-supervision and latent diffusion. arXiv preprint arXiv:2310.00318 .

[160] Sun, J., Wang, S., Zhang, J., Zong, C., 2019. Towards sentence-level brain decoding with distributed representations, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7047–7054.

[161] Sun, J., Wang, S., Zhang, J., Zong, C., 2020. Neural encoding and decoding with distributed sentence representations. IEEE Transactions on Neural Networks and Learning Systems 32, 589–603.

[162] Takagi, Y., Nishimoto, S., 2023a. High-resolution image reconstruction with latent diffusion models from human brain activity, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14453–14463. https://openaccess.thecvf.com/content/CVPR2023/html/Takagi_High-Resolution_Image_Reconstruction_With_Latent_Diffusion_Models_From_Human_Brain_CVPR_2023_paper.html.

[163] Takagi, Y., Nishimoto, S., 2023b. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. arXiv preprint arXiv:2306.11536 https://arxiv.org/abs/2306.11536.

[164] Tang, J., LeBel, A., Jain, S., Huth, A.G., 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. Nature Neuroscience 26, 858–866. https://www.nature.com/articles/s41593-023-01304-9.

[165] Tian, Z., Quan, R., Ma, F., Zhan, K., Yang, Y., 2025. Brainguard: Privacy-preserving multisubject image reconstructions from brain activities, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 14414–14422.

[166] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 https://arxiv.org/abs/2302.13971.

[167] Tripathy, K., Markow, Z.E., Fishell, A.K., Sherafati, A., Burns-Yocum, T.M., Schroeder, M.L., Svoboda, A.M., Eggebrecht, A.T., Anastasio, M.A., Schlaggar, B.L., et al., 2021. Decoding visual information from high-density diffuse optical tomography neuroimaging data. NeuroImage 226, 117516. https://www.sciencedirect.com/science/article/pii/S1053811920310016.

[168] Tripathy, K., Markow, Z.E., Fogarty, M., Schroeder, M.L., Svoboda, A.M., Eggebrecht, A.T., Schlaggar, B.L., Trobaugh, J.W., Culver, J.P., 2025. Multisensory naturalistic decoding with high-density diffuse optical tomography. Neurophotonics 12, 015002–015002. https://pubmed.ncbi.nlm.nih.gov/39850351/.

[169] Vandecappelle, S., Deckers, L., Das, N., Ansari, A.H., Bertrand, A., Francart, T., 2021. Eeg-based detection of the locus of auditory attention with convolutional neural networks. Elife 10, e56481. https://elifesciences.org/articles/56481.

[170] Värbu, K., Muhammad, N., Muhammad, Y., 2022. Past, present, and future of eeg-based bci applications. Sensors 22, 3331.

[171] Wang, C., Yan, H., Huang, W., Li, J., Wang, Y., Fan, Y.S., Sheng, W., Liu, T., Li, R., Chen, H., 2022a. Reconstructing rapid natural vision with fmri-conditional video generative adversarial network. Cerebral Cortex 32, 4502–4511. https://academic.oup.com/cercor/article/32/20/4502/6515038.

[172] Wang, J., Song, Z., Ma, Z., Qiu, X., Zhang, M., Zhang, Z., 2024. Enhancing eeg-to-text decoding through transferable representations from pre-trained contrastive eeg-text masked autoencoder. arXiv preprint arXiv:2402.17433 https://arxiv.org/abs/2402.17433.

[173] Wang, S., Zhang, J., Wang, H., Lin, N., Zong, C., 2020. Fine-grained neural decoding with distributed word representations. Information Sciences 507, 256–272.

[174] Wang, S., Zhang, X., Zhang, J., Zong, C., 2022b. A synchronized multimodal neuroimaging dataset for studying brain language processing. Scientific Data 9, 590.

[175] Wei, Z., Dan, T., Wu, G., 2025. Large connectome model: An fmri foundation model of brain connectomes empowered by brain-environment interaction in multitask learning landscape. arXiv preprint arXiv:2510.18910 .

[176] Wen, H., Shi, J., Zhang, Y., Lu, K.H., Cao, J., Liu, Z., 2018. Neural encoding and decoding with deep learning for dynamic natural vision. Cerebral cortex 28, 4136–4160. https://academic.oup.com/cercor/article/28/12/4136/4560155.

[177] Willett, F.R., Avansino, D.T., Hochberg, L.R., Henderson, J.M., Shenoy, K.V., 2021. High-performance brain-to-text communication via handwriting. Nature 593, 249–254. https://www.nature.com/articles/s41586-021-03506-2.

[178] Willett, F.R., Kunz, E.M., Fan, C., Avansino, D.T., Wilson, G.H., Choi, E.Y., Kamdar, F., Glasser, M.F., Hochberg, L.R., Druckmann, S., et al., 2023. A high-performance speech neuroprosthesis. Nature 620, 1031–1036. https://www.nature.com/articles/s41586-023-06377-x.

[179] Wong, D.D., Fuglsang, S.A., Hjortkjær, J., Ceolini, E., Slaney, M., De Cheveigne, A., 2018. A comparison of regularization methods in forward and backward models for auditory attention decoding. Frontiers in neuroscience 12, 531. https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2018.00531/full.

[180] Wu, H., Li, Q., Zhang, C., He, Z., Ying, X., 2025. Bridging the vision-brain gap with an uncertainty-aware blur prior, in: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 2246–2257.

[181] Xia, W., de Charette, R., Oztireli, C., Xue, J.H., 2024. Umbrae: Unified multimodal brain decoding, in: European Conference on Computer Vision, Springer. pp. 242–259. https://link.springer.com/chapter/10.1007/978-3-031-72667-5_14.

[182] Xiong, W., Ma, L., Li, H., 2025. Synthesizing intelligible utterances from eeg of imagined speech. Frontiers in Neuroscience 19, 1565848. https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2025.1565848/full.

[183] Xu, L., Xu, M., Jung, T.P., Ming, D., 2021. Review of brain encoding and decoding mechanisms for eeg-based brain–computer interface. Cognitive neurodynamics 15, 569–584.

[184] Xu, Y., Liu, B., Shao, W., Du, Y., He, S., Zhu, T., 2025. Cross-subject mind decoding from inaccurate representations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15066–15075.

[185] Xu, Z., Bai, Y., Zhao, R., Hu, H., Ni, G., Ming, D., 2022. Decoding selective auditory attention with eeg using a transformer model. Methods 204, 410–417. https://www.sciencedirect.com/science/article/pii/S1046202322000986.

[186] Yan, S., Fan, C., Zhang, H., Yang, X., Tao, J., Lv, Z., 2024. Darnet: dual attention refinement network with spatiotemporal construction for auditory attention detection, in: Proceedings of the 38th International Conference on Neural Information Processing Systems, pp. 31688–31707.

[187] Yin, C., Ye, Z., Li, P., 2024. Language reconstruction with brain predictive coding from fmri data. arXiv preprint arXiv:2405.11597 https://arxiv.org/abs/2405.11597.

[188] Yu, S., Chen, Z., Wu, W., Chen, S., Liu, Z., Nie, J., Xing, X., Xu, X., Zhang, X., 2025. From flat to round: Redefining brain decoding with surface-based fmri and cortex structure, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6189–6198.

[189] Zeng, B., Li, S., Liu, X., Gao, S., Jiang, X., Tang, X., Hu, Y., Liu, J., Zhang, B., 2024. Controllable mind visual diffusion model, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6935–6943. https://ojs.aaai.org/index.php/AAAI/article/view/28519.

[190] Zeng, H., Xia, N., Qian, D., Hattori, M., Wang, C., Kong, W., 2023. Dm-re2i: A framework based on diffusion model for the reconstruction from eeg to image. Biomedical Signal Processing and Control 86, 105125. https://www.sciencedirect.com/science/article/abs/pii/S174680942300558X.

[191] Zhang, H., Zhao, M., Wei, C., Mantini, D., Li, Z., Liu, Q., 2021. Eegdenoisenet: a benchmark dataset for deep learning solutions of eeg denoising. Journal of Neural Engineering 18, 056057. https://iopscience.iop.org/article/10.1088/1741-2552/ac2bf8/meta.

[192] Zhang, Y., Ruan, H., Yuan, Z., Du, H., Gao, X., Lu, J., 2023. A learnable spatial mapping for decoding the directional focus of auditory attention using eeg, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5. https://ieeexplore.ieee.org/abstract/document/10096819.

[193] Zhang, Y., Wang, S., Dong, X., Yu, J., Zong, C., 2024. Navigating brain language representations: A comparative analysis of neural language models and psychologically plausible models. arXiv preprint arXiv:2404.19364 .

[194] Zhang, Y., Wang, S., Lin, N., Dong, X., Li, C., Zong, C., 2025. Discovering semantic subdimensions through disentangled conceptual representations. Preprint .

[195] Zheng, R., Sun, L., 2024. Llm4brain: Training a large language model for brain video understanding. arXiv preprint arXiv:2409.17987 https://arxiv.org/abs/2409.17987.

[196] Zhou, J., Xie, Y., Fan, C., Wang, H., Lv, Z., Tao, L., 2025. Dhgcn: Dual hypergraph convolutional network for eeg-based auditory attention detection, in: Proceedings of the 33rd ACM International Conference on Multimedia, pp. 612–620. https://doi.org/10.1145/3746027.3755579.

[197] Zhou, M., Gong, Z., Dai, Y., Wen, Y., Liu, Y., Zhen, Z., 2023. A large-scale fmri dataset for human action recognition. Scientific Data 10, 415. https://www.nature.com/articles/s41597-023-02325-6.

[198] Zhou, M., Tian, C., Cao, R., Wang, B., Niu, Y., Hu, T., Guo, H., Xiang, J., 2018. Epileptic seizure detection based on eeg signals and cnn. Frontiers in neuroinformatics 12, 95. https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2018.00095/full.

[199] Zhu, D., Bieger, J., Garcia Molina, G., Aarts, R.M., 2010. A survey of stimulation methods used in ssvep-based bcis. Computational intelligence and neuroscience 2010, 702357. https://onlinelibrary.wiley.com/doi/full/10.1155/2010/702357.

[200] Zhu, S., Ye, Z., Ai, Q., Liu, Y., 2024. Eeg-imagenet: An electroencephalogram dataset and benchmarks with image visual stimuli of multi-granularity labels. arXiv preprint arXiv:2406.07151 https://arxiv.org/abs/2406.07151.

[201] Zhu, X., Chai, Y., Li, R., Lan, M., Gao, L., 2025. Decoding the moving mind: Multi-subject fmri-to-video retrieval with mllm semantic grounding. bioRxiv , 2025–04https://www.biorxiv.org/content/10.1101/2025.04.07.647335v1.abstract.

[202] Zink, R., Proesmans, S., Bertrand, A., Van Huffel, S., De Vos, M., 2017. Online detection of auditory attention with mobile eeg: closing the loop with neurofeedback. BioRxiv , 218727https://www.biorxiv.org/content/10.1101/218727v1.abstract.

[203] de Zuazo, X., Verbeni, V., Ku, L.C., Arrieta, E., Barrena, A., Klimovich-Gray, A., Saratxaga, I., Navas, E., Agirre, E., Molinaro, N., 2024. # neural2speech: Decoding speech and language from the human brain, in: Proc. IberSPEECH 2024, pp. 256–260. https://www.isca-archive.org/iberspeech_2024/dezuazo24b_iberspeech.html.