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

Yifan Wang[1], Shaonan Wang[2], Wenhao Cai[1], George Ford[1], Yang Cui[1], Yunhao Zhang[3,4], Changde Du[4,5], Cunhang Fan[6], Dongyang Li[7], Hongpeng Zhou[1], Hongyu Zhang[8], Jixing Li[9], Quanying Liu[7], Wei Huang[10], Yizhuo Lu[11,12], Zijiao Chen[13], Jingyuan Sun[1*]

[1*]Department of Computer Science, University of Manchester, Oxford Rd, Manchester, M13 9PL, Greater Manchester, United Kingdom.
[2]Department of Language Science and Technology, The Hong Kong Polytechnic University, 8 Hung Lok Road, Hung Hom, Kowloon, Hong Kong.
[3]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[4]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.
[5]State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[6]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, China.
[7]Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China.
[8]School of Control Science and Engineering, Shandong University, Jinan, China.
[9]Department of Linguistics and Translation, City University of Hong Kong, Tat Chee Avenue,Kowloon, Hong Kong.
[10]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China.

[11]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.
[12]School of Future Technology, University of Chinese Academy of Sciences, Beijing, China.
[13]School of Humanities and Sciences, Stanford University, 450 Jane Stanford Way, Stanford, 94305, CA, United States of America.

*Corresponding author(s). E-mail(s): jingyuan.sun@manchester.ac.uk;
Contributing authors: yifan.wang@manchester.ac.uk; shaonan.wang@polyu.edu.hk; wenhao.cai@postgrad.manchester.ac.uk; george.ford@postgrad.manchester.ac.uk; yang.cui@manchester.ac.uk; zhangyunhao2021@ia.ac.cn; changde.du@ia.ac.cn; cunhang.fan@ahu.edu.cn; lidy2023@mail.sustech.edu.cn; hongpeng.zhou@manchester.ac.uk; hongyu.zhang@mail.sdu.edu.cn; jixingli@cityu.edu.hk; liuqy@sustech.edu.cn; huangw@uestc.edu.cn; luyizhuo2023@ia.ac.cn; zijiao@stanford.edu;

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 Functional Magnetic Resonance Imaging (fMRI), Electroencephalography (EEG), and Magnetoencephalography (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 architectures—has expanded the feasible operating region of non-invasive brain decoding under controlled protocols, primarily through representation learning, neuro-semantic alignment, and strong generative priors. However, evidence for robust cross-subject and real-world deployment remains uneven, as many results are still demonstrated on limited cohorts or highly controlled settings. 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.

Keywords: Brain-Computer Interface (BCI), Non-Invasive Brain Decoding, Foundation Models, Deep Learning, Electroencephalography (EEG), Functional Magnetic Resonance Imaging (fMRI)

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 (da Silva 2013). 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 (Naselaris et al. 2009; Kay et al. 2008; Willett et al. 2021b). These achievements underscore the transformative possibilities, spanning healthcare innovations for neurological disorders, novel communication channels, and enhanced human-computer interaction paradigms (Pais-Vieira et al. 2013).

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) (da Silva 2013; Buzsáki et al. 2012; Logothetis 2008; Ferrari and Quaresima 2012). Unlike invasive methods, these techniques acquire neural signals from outside the body without surgical intervention, thereby avoiding associated health risks (Rossi et al. 2021). 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 (Abiri et al. 2019; Rubin et al. 2017).

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 (Radford et al. 2021; Gong et al. 2024). 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 (Radford et al. 2021; Devlin et al. 2019; Gong et al. 2024). Architectures such as the Transformer (Devlin et al. 2019) have enabled models like GPT (Radford et al. 2019) in language and CLIP (Radford et al. 2021) 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 (Gong et al. 2024).

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 (Värbu et al. 2022) or encoding/decoding (Xu et al. 2021; Abiri et al. 2019) often focus solely on specific modalities or emphasize algorithmic lists without unifying visual, linguistic, and multimodal fusion. Similarly, surveys on natural image reconstruction (Rakhimberdina et al. 2021) and fMRI decoding (Du et al. 2022) 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 establishes the theoretical foundations that motivate aligning computational models with biological neural processing, including neural coding, representational geometry, and related perspectives. Section 3 surveys the non-invasive decoding landscape by introducing major modalities and their constraints, and further summarizes commonly used datasets and benchmarks (Figure 1; Tables 1 and 2). Section 4 reviews core concepts and representative architectures of foundation models relevant to brain decoding (Table 3). Section 5 presents foundation-model–based decoding advances across vision, language/speech, and auditory attention/audio, guided by an overview timeline (Figure 2) and task-specific schematics (Figures 3 and 4), with datasets and reported evaluation metrics consolidated in Tables 4, 5 and a dedicated discussion on metric interpretation and compatibility. Section 6 synthesizes cross-domain trends and recurring methodological patterns. Section 7 discusses persisting challenges and future horizons, including validation regimes and reproducibility considerations, applicability boundaries contrasting FM-based and classical pipelines (Table 6), and a prioritized research roadmap (Table

7). Finally, Section 8 concludes with key takeaways and requirements for translating FM-driven non-invasive brain decoding from controlled demonstrations to reliable real-world applications.

Throughout this survey, we distinguish controlled feasibility demonstrations from translation-level evidence. In particular, claims about broad generalizability are considered well supported only when results are reproduced under cross-session and cross-subject evaluation, and ideally under naturalistic/online constraints.

2 Theoretical Foundations: Bridging Computational Models and Biological Neural Processing

To understand how foundation models successfully decode brain activity, it is imperative to explore the underlying computational neuroscience theories. This section elucidates the mathematical and biological principles that justify the alignment between artificial architectures and neural substrates.

2.1 The Neural Coding Hypothesis and Decodability

The fundamental premise of brain decoding is the **Neural Coding Hypothesis**, which posits that mental states including perceptions, intentions, and semantic thoughts are not abstract entities but are represented through specific physical patterns of neural activity.

Mathematically, we define a mapping function f such that a stimulus or internal state $\mathbf{s}(t)$ yields a neural response $\mathbf{r}(t)$:

$$\mathbf{r} = f(\mathbf{s}) \tag{1}$$

Decoding is feasible only if this mapping is systematic rather than stochastic. According to the Universal Approximation Theorem (Hornik et al. 1989), a neural network with sufficient capacity can approximate any continuous function. Thus, if the encoding process f is information-preserving, there exists an inverse mapping g such that:

$$\hat{\mathbf{s}} = g(\mathbf{r}) \approx f^{-1}(\mathbf{r}) \tag{2}$$

Beyond representational capacity, the Statistical Identifiability of the neural signal is crucial. For distinct mental states $\mathbf{s}_1 \neq \mathbf{s}_2$, the conditional distributions must satisfy $p(\mathbf{r}|\mathbf{s}_1) \neq p(\mathbf{r}|\mathbf{s}_2)$. This discriminability forms the basis of Bayesian optimal decoding, where the model seeks to approximate the posterior $\arg\max_{\mathbf{s}} p(\mathbf{s}|\mathbf{r})$ (Paninski et al. 2007).

2.2 Representational Geometry and Isomorphism

Recent success in utilizing Large Language Models (LLMs) for decoding stems from the theory of Geometric Isomorphism. This theory suggests that the relational structure of neural representations mirrors the structure of semantic embeddings:

$$d_{neural}(\mathbf{s}_i, \mathbf{s}_j) \approx d_{semantic}(\mathbf{s}_i, \mathbf{s}_j) \tag{3}$$

When the distance metrics between neural manifolds and artificial latent spaces are preserved, simple linear transformations (e.g., ridge regression) suffice to align the two domains (Kriegeskorte et al. 2008). This isomorphism explains why high-dimensional semantic embeddings from foundation models can serve as effective proxies for biological neural representations (Haxby et al. 2001).

2.3 Key Theoretical Perspectives on Neural Processing

We categorize the current computational understanding of brain-model alignment into three primary schools of thought:

- **Bayesian and Predictive Coding:** Proponents of this view, such as (Friston 2009) and (Bottemanne 2025), argue that the brain functions as a hierarchical generative model. It minimizes "surprisal" by constantly updating internal beliefs through probability inference. In this context, decoding is the process of extracting the brain's internal "prior" and "prediction error" signals, which can be mapped to synaptic plasticity and Bayesian updates.
- **Traditional Neural Coding:** This school focuses on the structural and temporal properties of neuronal population activity. Works by (Olshausen and Field 2004) emphasize *sparse coding*, while (Tsien 2016) introduces the "Power-of-N" combinatorial logic. Notably, Tsien suggests that "rare silence" (synchronized inactive states) within neuronal cliques may carry primary information, providing a new perspective on neural variability and self-information encoding that transcends simple rate coding.
- **Goal-Driven Artificial Neural Networks (ANNs):** This perspective, championed by (Yamins and DiCarlo 2016) and (Eickenberg et al. 2017), treats deep networks as functional surrogates for the brain. It posits that optimizing a model on a complex task (e.g., object recognition or next-token prediction) naturally leads to internal representations that are hierarchical and functionally similar to those found in the human cortex, even if the biological implementation (spikes vs. gradients) differs.

3 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).

3.1 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 (da Silva 2013; Buzsáki et al. 2012). Its chief advantage is millisecond-level temporal resolution, which enables tracking rapid neural dynamics and supports closed-loop or real-time applications (Zhu et al. 2024). 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 (da Silva 2013); recordings are also vulnerable to physiological artifacts (eye blinks, muscle activity) and environmental electrical noise (Zhang et al. 2021).

Magnetoencephalography (MEG) measures the minute magnetic fields produced by the same neural currents that underlie EEG signals (da Silva 2013; Buzsáki et al. 2012). 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 (da Silva 2013). 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 (Logothetis 2008). 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 (Logothetis 2008; Naselaris et al. 2009; Kay et al. 2008). Its principal limitation is temporal: the hemodynamic response peaks several seconds after neural events, resulting in second-level temporal resolution (Logothetis 2008). 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 (Ferrari and Quaresima 2012). 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 (Ferrari and Quaresima 2012). However, limited penetration depth restricts sensitivity primarily to superficial cortical tissue, and spatial precision and coverage remain below those of fMRI (Ferrari and Quaresima 2012).

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 (Ferrari and Quaresima 2012). 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 (da Silva 2013; Logothetis 2008; Ferrari and Quaresima 2012). 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.

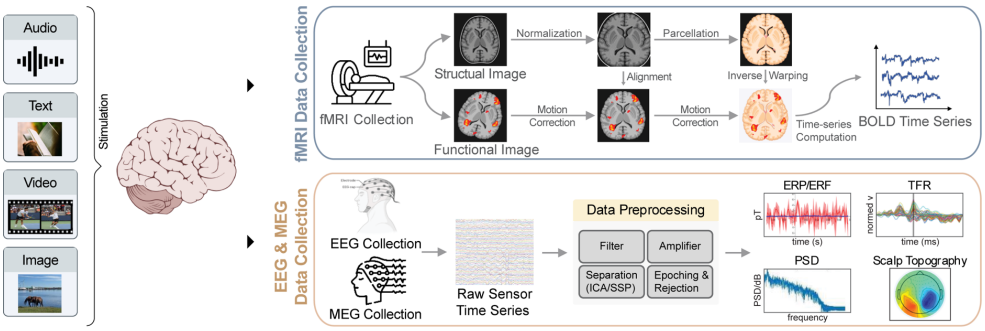

Fig. 1. Non-invasive data collection and preprocessing across fMRI, EEG and MEG. The diagram highlights typical stimuli, acquisition hardware, core preprocessing steps (e.g., normalization/alignment/motion correction for fMRI; filtering/ICA/epoching for EEG/MEG), and the resulting analysis-ready outputs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

3.2 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 (Abiri et al. 2019; Luo et al. 2024; Geirnaert et al. 2021; Shirakawa et al. 2024). 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 (Zhang et al. 2021; Geirnaert et al. 2021). 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 (Luo et al. 2024; Chen et al. 2023a). Third, hard physical limits in spatial and temporal resolution cap extractable information: slow

Table 1. Comparison of Non-Invasive BCI Modalities.

Modality	**Principle**	**Spatial Resolution**	**Temporal Resolution**	**Invasiveness**	**Portability**	**Cost**	**Key Strengths**	**Key Weak-nesses**
EEG	Electrical potentials from neural currents	Low (cm)	High (ms)	Non-invasive	High	Low	High temporal resolution, portable, low cost	Low spatial resolution, susceptible to artifact-s/noise
MEG	Magnetic fields from neural currents	Moderate (mm–cm)	High (ms)	Non-invasive	Low (requires shielding)	Very High	High temporal resolution, better spatial resolution 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 resolution, indirect measure, expensive, immobile, noisy
fNIRS	Light absorption by blood hemoglobin	Moderate (cm)	Moderate (seconds)	Non-invasive	Moderate–High	Moderate	Portable, balanced trade-off, movement tolerant	Lower spatial resolution than fMRI, limited depth penetration

hemodynamics blur rapid events in fMRI (Logothetis 2008), while volume conduction in EEG mixes sources and degrades spatial separability (da Silva 2013).

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 (Zink et al. 2017). Ecological validity is another barrier: models tuned on tightly controlled laboratory paradigms (e.g., P300 (Pritchard 1981; Solon et al. 2019), SSVEP (Zhu et al. 2010), or motor imagery (Altaheri et al. 2023; McGeady et al. 2019)) often degrade in naturalistic settings with unconstrained behavior, rich sensory input, and movement, creating a persistent "lab-to-life" gap (Luo et al. 2024; Abiri et al. 2019). 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 (Jiang et al. 2024; Shirakawa et al. 2024).

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.

While we have discussed accidental artifacts such as those caused by ocular, muscular, or environmental interference, the real-world deployment of BCI technology also introduces the potential for intentional corruption. Recent studies highlight adversarial perturbations; small, structured injections of noise which are optimized to mislead a trained model and induce large decoding errors (Zhang et al. 2020; Lubars and Chin 2022; Li et al. 2021). As the field shifts towards large foundation models, these threats to data integrity extend from test-time evasion to backdoor attacks, where triggers are implanted during pre-training, enabling specific mis-decoding to occur when a trigger pattern, such as an adversarial perturbation, is present (Meng et al. 2022). This distinction matters because robustness to incidental artifacts/'noise' does not imply robustness to intentional perturbations, which have been shown to bypass standard denoising pipelines remaining undetected (S-a et al. 2020).

3.3 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) (De Taillez et al. 2020; Wong et al. 2018) 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) (Wong et al. 2018; Fiedler et al. 2017; De Cheveigné et al. 2018; Dmochowski et al. 2018) showed declining performance in short time windows crucial for real-time use (Cai et al. 2020; Vandecappelle et al. 2021).

Deep learning (CNNs, RNNs) offered improvements by automatically learning features and temporal dependencies (Craik et al. 2019; Zhou et al. 2018), enhancing

tasks like motor imagery classification (Altaheri et al. 2023), ERP detection (Solon et al. 2019), fine-grained semantic decoding (Wang et al. 2020), and even early visual reconstruction (Nishimoto et al. 2011; Beliy et al. 2019; Shen et al. 2019). However, these models, often trained from scratch on limited BCI datasets, still struggled with generalization due to data scarcity and variability (Luo et al. 2024), had difficulty decoding complex semantic content and often required extensive preprocessing (Craik et al. 2019).

3.4 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 (Zhang et al. 2021; He et al. 2022; Chen et al. 2023a). 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 (Luo et al. 2024; Tang et al. 2023). 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 (Radford et al. 2021; Mitchell et al. 2008; Huth et al. 2016; Gong et al. 2024). 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 (Scotti et al. 2024; Liu et al. 2023; Tang et al. 2023). 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 (Takagi and Nishimoto 2023a; Radford et al. 2019; Gong et al. 2024).

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

(Jiang et al. 2024; Gong et al. 2024; Jayalath et al. 2024). 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.

3.5 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.

3.5.1 Image Stimulus

Image datasets are fundamental to visual decoding research, primarily focusing on static scene and object representations. Typical datasets include NSD (Allen et al. 2022) and GOD (Horikawa and Kamitani 2017), which are used for high-resolution brain imaging of natural scenes and category recognition, respectively; EEG-AVC (Spampinato et al. 2017) and Things-EEG2 (Gifford et al. 2022) utilize high-temporal resolution EEG to record rapid visual categorization. Things-MEG (Hebart et al. 2023) 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.

3.5.2 Video Stimulus

Video datasets emphasize temporal information and ecological validity. These include DNV (Wen et al. 2018), STN (Seeliger et al. 2019), and HAD (Zhou et al. 2023) in fMRI, which are used for natural scene or action recognition tasks. In EEG, SEED-DV (Liu et al. 2024b) combines rich semantic annotations with multimodal signals for semantic decoding from EEG to video. MEG Narrative (Armeni et al. 2022) and HD-DOT (Tripathy et al. 2021) 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.

3.5.3 Audio Stimulus

Auditory and speech datasets typically focus on natural narratives. For example, CLSR (Tang et al. 2023), NL fMRI (LeBel et al. 2023), and Narratives fMRI (Nastase et al. 2021) use fMRI to record brain activity while participants listen to natural stories. The Narratives fMRI dataset (Nastase et al. 2021) is large (345 subjects), making it a

Table 2. **Summary of datasets.** Summary of datasets grouped by imaging modality and stimulus type used or referenced in this work.

Modality	Stimulus	Dataset	Task / Stimuli	Paradigm	Subjects
fMRI	Image	NSD (Allen et al. 2022)	Natural Scene Pictures	Natural Vision	8
		GOD (Horikawa and Kamitani 2017)	Object Category Images	Category Identification / Representation	5
	Video	DNV (Wen et al. 2018)	Natural Video Viewing	Naturalistic Video	3
		STN (Seeliger et al. 2019)	Natural Video Viewing	Naturalistic Video	1
		HAD (Zhou et al. 2023)	Human Action Recognition	Category Identification / Representation	30
	Audio	CLSR (Tang et al. 2023)	Naturalistic Narrative	Naturalistic Narrative	7
		NL fMRI (LeBel et al. 2023)	Naturalistic Narrative	Naturalistic Narrative	8
		Narratives (Nastase et al. 2021)	Naturalistic Narrative	Naturalistic Narrative	345
	Multi Stimuli	HCP (Elam et al. 2021)	Multi Task	Multi-paradigm Resource	1200
	Reading/Audio	MOUS (Schoffelen et al. 2019)	Language Processing	Multi-paradigm Resource	204
EEG	Image	EEG-AVC (Spampinato et al. 2017)	Visual Object Classification	ERP	6
		Things-EEG2 (Gifford et al. 2022)	Fast Sequence of Natural Images	ERP	10
	Video	SEED-DV (Liu et al. 2024b)	Natural Video Viewing	Naturalistic Video	20
		MOABB (Jayaram and Barachant 2018)	Multi Task	MI / ERP / SSVEP	-
	Reading	ZuCo (Hollenstein et al. 2018)	Reading / Language Comprehension	ERP	12
	Reading/Img	ISD (Xiong et al. 2025)	Imagine Pronunciation	Speech Imagery	13
MEG	Image	Things-MEG (Hebart et al. 2023)	Natural Scene Pictures	Category Identification / Representation	4
	Multi Stimuli	HCP (Elam et al. 2021)	Multi Task	Multi-paradigm Resource	95
	Reading/Audio	MOUS (Schoffelen et al. 2019)	Language Processing	Multi-paradigm Resource	204
	Video	MEG Narrative (Armeni et al. 2022)	Naturalistic Narrative	Naturalistic Narrative	3
HD-DOT	Video	HD-DOT (Tripathy et al. 2021)	Natural Video Viewing	Naturalistic Video	3

landmark resource for natural language decoding. The EEG Imagined Speech Dataset (ISD) (Xiong et al. 2025) covers both real and imagined speech tasks, supporting end-to-end research on direct speech generation from brain signals.

3.5.4 Reading/Language Stimulus

Reading and language datasets focus on the neural mechanisms of language processing. ZuCo (Hollenstein et al. 2018) combines EEG and eye tracking to record brain activity while participants read natural sentences. MOUS (Schoffelen et al. 2019) and Narratives fMRI (Nastase et al. 2021) cover reading and auditory language tasks, and provide structural and diffusion MRI for multimodal fusion. ISD (Xiong et al. 2025) involves speech imagery and language production, and is suitable for brain decoding tasks related to language production.

3.5.5 Multi-stimuli

Some datasets cover a variety of stimulus and task types, supporting multi-paradigm comparisons and cross-modal studies. For example, both HCP (Elam et al. 2021) and MOUS (Schoffelen et al. 2019) 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 (Jayaram and Barachant 2018) integrates multiple public EEG datasets and standardizes paradigms such as MI, ERP, and SSVEP, making it a valuable resource for algorithm comparison and benchmarking.

4 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.

4.1 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 (Radford et al. 2021; Devlin et al. 2019). 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 (Kaplan et al. 2020; Henighan et al. 2020; Hestness et al. 2017). Pre-training objectives instantiate this breadth: masked-token prediction in language (e.g., BERT) and

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

cross-modal contrastive alignment in vision–language (e.g., CLIP) endow models with semantic structure, syntactic competence, and cross-modal correspondences without manual annotation (Devlin et al. 2019; Radford et al. 2021).

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 (Tang et al. 2023).

Prominent exemplars include large language models (GPT (Radford et al. 2019), BERT (Devlin et al. 2019)), vision–language encoders (CLIP (Radford et al. 2021)), and modern generative models (e.g., diffusion backbones used for image synthesis and reconstruction (Takagi and Nishimoto 2023a)). Conceptually, the paradigm shifts from crafting bespoke pipelines to leveraging a powerful, pre-trained base that can be aligned with specific BCI objectives through fine-tuning or prompting.

4.2 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 (Radford et al. 2021, 2019); 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 (Mitchell et al. 2008; Radford et al. 2021; Liu et al. 2023; Gong et al. 2024). 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 (Zhang et al. 2025). Notably, larger language models often exhibit stronger brain alignment, suggesting scaling benefits for representational fidelity (Antonello et al. 2023; Gao et al. 2025a). 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 (Zhang et al. 2025).

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 (Luo et al. 2024; Tang et al. 2023; Li et al. 2024). 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 (Takagi and Nishimoto 2023a) backbones and autoregressive LLMs (Radford et al. 2019) 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 (Chen et al. 2023b; Sun et al. 2023a; Chen et al. 2023a; Scotti et al. 2024). 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 (Gong et al. 2024).

4.3 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 (Fan et al. 2025b,a; Li et al. 2025b; Fan et al. 2024). 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 (Devlin et al. 2019). 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 (Radford et al. 2019; Touvron et al. 2023).

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 (Dosovitskiy et al. 2020; He et al. 2022). 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 (Radford et al. 2021). For generative targets, diffusion backbones now dominate high-fidelity reconstructions from brain activity, superseding earlier GAN-based (Goodfellow et al. 2020) 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 (Baevski et al. 2020). 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 (Qiu et al. 2025b).

5 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,

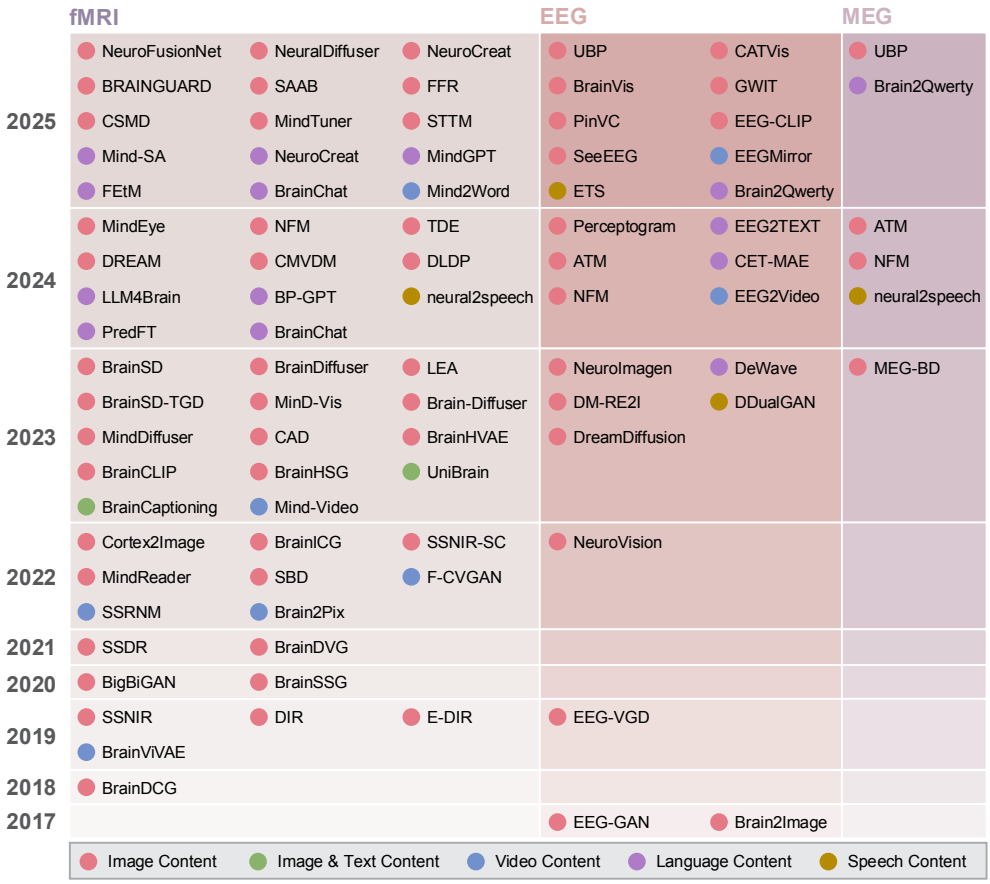

Fig. 2. Chronological map of representative non-invasive brain-decoding works (2017–2025) across modalities (fMRI/EEG/MEG). Colors denote output content types (image, video, language, speech). This overview situates subsequent method sections and helps cross-reference modality-specific trends.

speech, and auditory attention. A common conceptual pipeline involves: 1) Extracting denoised representations from neural signals (often using MAEs or SSL (Zhang et al. 2021; Chen et al. 2023a; Jiang et al. 2024; Jayalath et al. 2024)), 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 (Liu et al. 2023; Tang et al. 2023; Li et al. 2024; Chen et al. 2024b), and 3) Generating decoding results using conditional FMs (Diffusion Models, LLMs) guided by the aligned neural representations, leveraging strong priors for high-quality output (Chen et al. 2023b; Sun et al. 2023a; Chen et al. 2023a; Scotti et al. 2024). 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.

5.1 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 (He et al. 2022), wav2vec 2.0 (Baevski et al. 2020), and BLIP/CLIP (Li et al. 2022), providing a robust foundation for small sample sizes and cross-subject learning (Radford et al. 2021; Rombach et al. 2022). 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 (Liu et al. 2023; Radford et al. 2021; Rombach et al. 2022). Third, they leverage strong generative priors (diffusion/autoregression) to transform weak and time-delayed neural conditions into readable image, text, and speech outputs (Rombach et al. 2022; Takagi and Nishimoto 2023a; Ozcelik and VanRullen 2023; Lu et al. 2023). 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 (Tang et al. 2023; Liu et al. 2023). Meanwhile, native neuroimaging models (such as M-DBPNet (Fan et al. 2025b), ListenNet (Fan et al. 2025a), MHANet (Li et al. 2025b), 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 (Shehzad et al. 2024; Kim et al. 2023). 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 (Wei et al. 2025). 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 (Chen et al. 2023a). This paradigm shift moves neuroimaging analysis from fixed, statistical connectivity measures towards flexible, learnable, and generalizable neuro-relational embeddings.

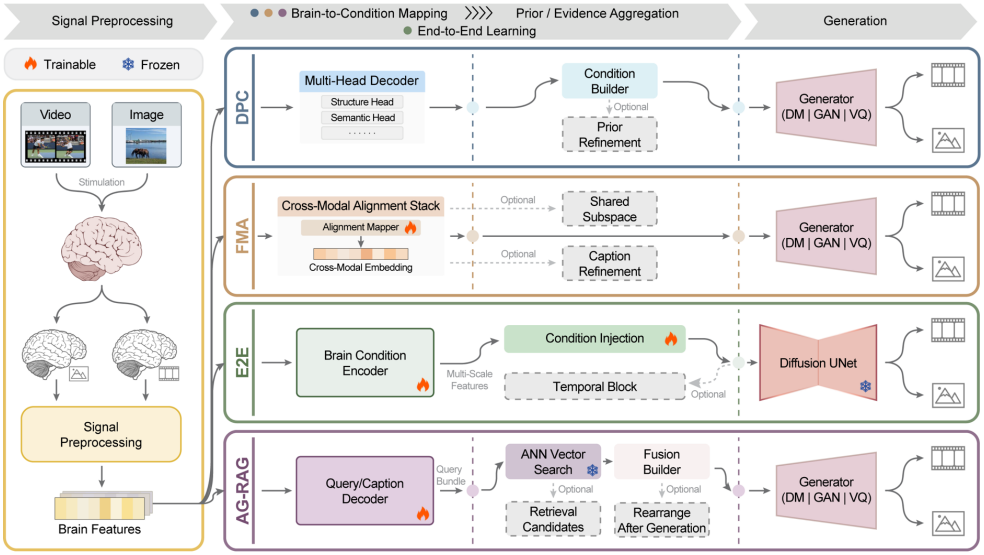

Fig. 3. Decoding Visual Experiences Flowchart. This flowchart shows the different types of methods for images and video decoding tasks. (1)DPC (Decoded-Prior Conditioning): Multi-head decoders predict structural (VAE latent, depth/edges) and semantic (CLIP/caption) priors, fused and fed to a generator (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.

5.2 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 (Liu et al. 2025a). The dominant strategy combines CLIP's representational power with diffusion models' generative prowess (Gong et al. 2024). An encoder maps fMRI signals to CLIP's shared embedding space, capturing semantics (Gong et al. 2024; Liu et al. 2023). These embeddings condition a latent diffusion model (e.g., Stable Diffusion (Takagi and Nishimoto 2023a)) to generate the image (Scotti et al. 2024; Gong et al. 2024; Sun et al. 2023b). CLIP provides semantic guidance, while diffusion models ensure

fidelity (Chen et al. 2023a; Gong et al. 2024). 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 (Chen et al. 2023a). Brain-Diffuser predicted multimodal CLIP features from fMRI to condition Versatile Diffusion. These FM-based methods yield significant improvements in semantic accuracy and perceptual fidelity over earlier approaches (Chen et al. 2023a; Liu et al. 2025a).

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 (Li et al. 2024; Craik et al. 2019; Bai et al. 2023; Song et al. 2023). 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 (Benchetrit et al. 2023).

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 (Gong et al. 2024; Chen et al. 2023b; Zhu et al. 2025; Sun et al. 2024).

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 (Liu et al. 2025a; Craik et al. 2019; Li et al. 2024). At the same time, the degree to which this "neural-to-semantic alignment + conditional diffusion generation" recipe generalizes beyond controlled benchmarks and subject-specific settings is still inconsistently demonstrated.

Failure modes & boundary conditions. FM-conditioned reconstruction often depends on accurate subject-specific mapping into a shared semantic latent, which may degrade under session drift, noise, or mismatched preprocessing. Diffusion priors can yield visually plausible images even when neural evidence is weak, creating a perceptual-quality vs neural-fidelity gap. Cross-dataset transfer is often fragile because voxel/channel selection, stimulus distributions, and evaluation pipelines differ.

Contradictions across studies. Reported gains are not always comparable because papers optimize different targets and use different conditioning signals and sampling settings. Methods emphasizing semantic alignment may trade off fine-grained structural fidelity, and "State-of-the-art" claims can be confounded by within-subject vs cross-subject protocols. Standardized metrics and splits are therefore necessary to interpret improvements consistently.

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 (Follows our taxonomy (DPC/FMA/E2E/AG-RAG/AG-EML)), Metrics (Perceptual Similarity Metrics (PSM) include CLIP, C3D (Convolutional 3D), AlexNet, Inception, EffNet-B, SwAV). FID is Fréchet Inception Distance. mIoU is Mean Intersection over Union. PCC is Pearson Correlation Coefficient. CD is Chamfer Distance, EMD is Earth Mover's Distance. IS is Inception Score. IC is Inception Classification Accuracy.

Brain Modality	Decoded Content	Dataset	Representation Model	Generation Model	Work (Reference)	Method	Metrics
fMRI	Image	NSD (Allen et al. 2022)	CLIP	SDM	BrainSD (Takagi and Nishimoto 2023a)	DPC	PSM, Human Acc.
			CLIP	SDM	BrainSD-TGD (Takagi and Nishimoto 2023b)	DPC	PSM
			VQ-VAE	SDM	MindDiffuser (Lu et al. 2023)	E2E	PSM, SSIM, FID
			CLIP	-	BrainCLIP (Liu et al. 2023)	FMA	Recall@K, PSM
			CLIP	LDM	BrainDiffuser (Ozcelik and VanRullen 2023)	DPC	PSM, PixCorr
			CLIP	DALL-E2	MindEye (Scotti et al. 2024)	FMA	PSM, PixCorr, SSIM, Retrieval Acc.
			CLIP	GAN	MindReader (Lin et al. 2022)	AG-RAG	FID, Retrieval Acc., AUC-ROC
			CLIP	LDM	DLDP (Ferrante et al. 2024b)	AG-RAG	PSM, SSIM, Retrieval Acc.
			CLIP	-	NeuroFusionNet (Lang et al. 2025)	DPC	mIoU, PCC
			CLIP	DM	NeuralDiffuser (Li et al. 2025a)	FMA	PSM, PixCorr, SSIM, Retrieval Acc.
			CLIP	DM	BRAINGUARD (Tian et al. 2025)	FMA	PSM, PixCorr, SSIM
			CLIP	DM	MindTuner (Gong et al. 2025)	FMA	PSM, PixCorr, SSIM, Retrieval Acc.
			CLIP	DM	CSMD (Xu et al. 2025)	FMA	PSM, PixCorr, SSIM

Continued on next page...

Table 4 – Continued from previous page

Brain Modality	Decoded Content	Dataset	Representation Model	Generation Model	Work (Reference)	Method	Metrics
fMRI	Image	NSD (Allen et al. 2022)	CLIP	DM	FFR (Yu et al. 2025)	FMA	PSM, PixCorr, SSIM
			CLIP	-	NFM (Ferrante et al. 2024c)	FMA	PSM, Retrieval Acc.
		GOD (Horikawa and Kamitani 2017)	-	DM	CMVDM (Zeng et al. 2024)	AG-RAG	PCC, SSIM, N-way Acc.
			SC-MBM	DC-LDM	MinD-Vis (Chen et al. 2023a)	E2E	FID, N-way Acc.
			DC-MAE	LDM	CAD (Sun et al. 2023a)	DPC	N-way Acc.
			CNN	LDM	SBD (Ferrante et al. 2022)	AG-RAG	FID, Wu-Palmer Similarity, Human Acc.
			D-VAE	GAN	BrainDVG (Ren et al. 2021)	DPC	Linear Correlation, SSIM, N-way Acc.
			-	GAN	BigBiGAN (Mozafari et al. 2020)	DPC	PCC, MSE
			SwAV ResNet-50	IC-GAN	BrainICG (Ozcelik et al. 2022)	DPC	PSM, PixComp, SSIM
			VDVAE	LDM	Brain-Diffuser (Meng and Yang 2023)	DPC	PSM, PixCorr, SSIM
			CLIP	MaskGIT	LEA (Qian et al. 2023)	FMA	PSM, FID, N-way Acc.
			CLIP	DM	SAAB (Kalantari et al. 2025)	FMA	PSM, SSIM, Identification Acc.
		NSD&GOD (Horikawa and Kamitani 2017; Allen et al. 2022)	CLIP	SDXL-Turbo	NeuroCreat (Jing et al. 2025a)	FMA	PSM, PixCorr, SSIM
			CLIP	DM	STTM (Liu et al. 2025c)	FMA	PSM, PixCorr, SSIM, Retrieval Acc.
		ImageNet (Deng et al. 2009)	VGG19	LDM	(Kalantari et al. 2025)	DPC	PSM, SSIM, Identification Acc.
	Image & Text	NSD (Allen et al. 2022)	CLIP	LDM	UniBrain (Mai and Zhang 2023)	DPC	PSM, PixCorr, SSIM, Meteor, Rouge-1, Rouge-L
			CLIP	DM	Brain Captioning (Ferrante et al. 2023)	DPC	PSM, PixCorr, SSIM
	Video	DNV (Wen et al. 2018)	-	SDM	Mind-Video (Chen et al. 2023b)	AG-RAG	SSIM, N-way Acc.

Continued on next page...

Table 4 – Continued from previous page

Brain Modality	Decoded Content	Dataset	Representation Model	Generation Model	Work (Reference)	Method	Metrics
fMRI	Video	DNV (Wen et al. 2018)	MARS	De-CNN	SSRNM (Kupershmidt et al. 2022)	DPC	PSM, PCC, SSIM
			VAE(Encoder)	VAE(Decoder)	BrainViVAE (Han et al. 2019)	DPC	SSIM, Circular Correlation
			-	fMRI-GAN	f-CVGAN (Wang et al. 2022a)	DPC	SSIM, PSNR, MSE
			CLIP	DM	Mind2Word (Jiang et al. 2025)	FMA	SSIM, N-way Acc.
		STN (Seeliger et al. 2019)	-	FCN	Brain2Pix (Le et al. 2022)	DPC	PSM, PCC, Euclidean Dis.
	3D	Self Dataset (Gao et al. 2025b)	CLIP	DM	MinD-3D++ (Gao et al. 2025b)	FMA	LPIPS, PSNR, SSIM, N-way Acc., CD, EMD
EEG	Image	EEG-AVC (Spampinato et al. 2017)	CLIP	DM	NeuroImagen (Lan et al. 2023)	DPC	IS, SSIM, N-way Acc.
			EVRNet	DM	DM-RE2I (Zeng et al. 2023)	E2E	Acc., IS
			Visual-Guided CNN	GAN	EEG-VGD (Jiao et al. 2019)	DPC	Classification Acc.
			-	cGAN	EEG-GAN (Palazzo et al. 2017)	AG-RAG	IS, IC
			-	GAN	NeuroVision (Khare et al. 2022)	DPC	Acc., IS
			CLIP	DM	BrainVis (Fu et al. 2025a)	FMA	IS, FID, N-way Acc.
			-	DM	GWIT (Lopez et al. 2025)	FMA	IS, FID, LPIPS, N-way Acc.
		Things-EEG2 (Gifford et al. 2022)	CLIP	LDM	Perceptogram (Fei et al. 2024)	DPC	PSM, PixCorr, SSIM
			CLIP	DM	ATM (Li et al. 2024)	E2E	PSM, PixCorr, SSIM
			CLIP	-	UBP (Wu et al. 2025)	FMA	N-way Acc.
			CLIP	SDXL-Turbo	EEG-CLIP (Cao et al. 2025)	FMA	PSM, PixCorr, SSIM
			CLIP	DM	SeeEEG (Kim et al. 2025)	FMA	PSM, PixCorr, SSIM, N-way Acc.
		Self-collected Dataset (Guo et al. 2023)	LSTM Network	DCGAN	Brain2Image (Kavasidis et al. 2017)	DPC	IS, IC

Continued on next page...

Table 4 – Continued from previous page

Brain Modality	Decoded Content	Dataset	Representation Model	Generation Model	Work (Reference)	Method	Metrics
EEG	Image	MOABB (Jayaram and Barachant 2018)	CLIP	DM	DreamDiffusion (Bai et al. 2023)	E2E	Acc.
		ImageNet (Deng et al. 2009)	CNN	LDM	(Ferrante et al. 2024a)	DPC	Kappa, N-way Acc.
			CLIP	DM	CATVis (Mehmood et al. 2025)	FMA	FID, IS, N-way Acc.
		NSD (Allen et al. 2022)	CLIP	-	NFM (Ferrante et al. 2024c)	FMA	PSM, N-way Acc.
	Video	SEED-DV (Liu et al. 2024b)	CLIP	DM	EEGMirror (Liu et al. 2025b)	FMA	SSIM, N-way Acc.
MEG	Image	Things-MEG (Hebart et al. 2023)	DINOv2, CLIP	LDM	MEG-BD (Benchetrit et al. 2023)	FMA	PSM, PixCorr, SSIM
			CLIP	-	UBP (Wu et al. 2025)	FMA	N-way Acc.
		Things-EEG2 (Gifford et al. 2022)	CLIP	DM	ATM (Li et al. 2024)	E2E	PSM, PixCorr, SSIM
		NSD (Allen et al. 2022)	CLIP	-	NFM (Ferrante et al. 2024c)	FMA	PSM, N-way Acc.

Reliability & reproducibility signals. Evidence is concentrated on a few canonical benchmarks (e.g., NSD/GOD summarized in Table 4), which aids comparability but risks overfitting to shared routines. Reliability is stronger when multi-subject and cross-session results are reported with clear pre-processing and split details, yet these signals are uneven. We recommend routine reporting of subject/session counts, split strategies, and both fidelity- and semantics-oriented metrics.

5.3 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) (Sato and Kobayashi 2025; Chen et al. 2024b; Qiu et al. 2025b). 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 (Sato and Kobayashi 2025). A common recipe builds an encoding model that predicts fMRI responses from linguistic features (Sun et al. 2020) and then scores candidate continuations sampled from an LLM (e.g., GPT), selecting the sequence that best

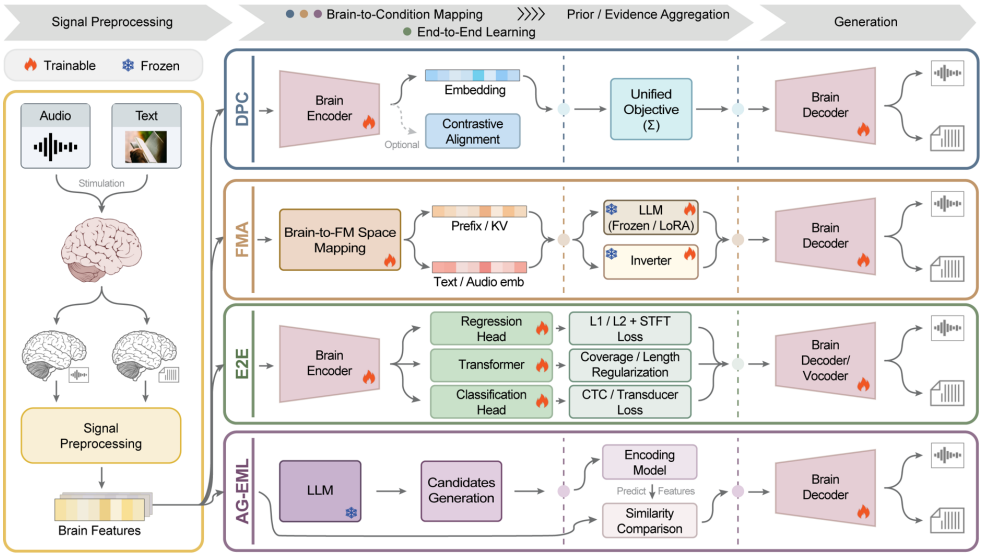

Fig. 4. Decoding Language and Speech Flowchart. This flowchart shows the different types of methods for language and speech decoding tasks. (1) DPC (Decoded-Prior Conditioning): Preprocessed brain signals are encoded by a Brain Encoder and decoded by a Unified Decoder (text/speech) under a Regularized 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.

matches the predicted brain activity—trading data for fluency and substantially mitigating temporal limits (Tang et al. 2023). 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 (Zhang et al. 2024). 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 (Tang et al. 2023; Antonello et al. 2023). For example: *MindLLM* targets subject-agnostic decoding via a neuroscience-informed attention

encoder plus "Brain Instruction Tuning" with an off-the-shelf LLM (Qiu et al. 2025a); *UMBRAE* aligns fMRI into a VLM latent space to flexibly decode text and/or images (Xia et al. 2024); and *GPT-2 prompting* directly treats fMRI as prompts aligned to text prompts through contrastive learning (Chen et al. 2024b). Early language decoding work primarily focused on lexical-level decoding. Subsequently, research began to shift towards more complex sentence-level semantic representations (Sun et al. 2019). Overall, FMs supply powerful linguistic constraints and generation, advancing versatility and cross-subject generalization while compensating for fMRI's temporal limitations (Sato and Kobayashi 2025; Tang et al. 2023; Qiu et al. 2025a).

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 (Craik et al. 2019; Qiu et al. 2025b). 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 (Jayalath et al. 2024). Aligning MEG features to pretrained speech models further boosts recognition of perceived speech and supports time-resolved probing of linguistic feature dynamics (Jayalath et al. 2024; Qiu et al. 2025b).

Looking toward practical communication BCIs, non-invasive routes based on fMRI (Tang et al. 2023), potentially fNIRS, and EEG spellers augmented with FM priors (Carìa 2025) show growing promise but face constraints in equipment bulk, calibration demands, latency, and robustness. Across modalities, the FM recipe—neural-to-semantic/acoustic alignment plus generative or retrieval decoding—often enables richer outputs from weak signals, yet robust cross-session and cross-subject performance remains unevenly validated.

Failure modes & boundary conditions. Language/speech pipelines are sensitive to intermediate targets (tokens vs embeddings vs acoustics) and to the quality of brain-to-linguistic mapping. Injecting brain-derived conditions into frozen LLMs can be brittle when the condition signal is noisy, especially for open-vocabulary generation, while end-to-end routes often require more paired data and careful regularization. For imagined-speech EEG, low SNR and inter-session variability frequently dominate performance.

Contradictions across studies. Results diverge because evaluation emphasizes different endpoints (WER/CER, semantic similarity or perceived intelligibility) and different constraints (closed vocabulary/spellers and open-ended generation). Protocols also vary in whether candidate reranking (AG-EML) is used, which changes task difficulty. Without standardized splits and explicit decoding constraints, cross-paper ranking remains partially confounded.

Table 5. Summary Table of Recent Language/Speech Decoding Works. To help compare the upper limits of language/speech decoding, a small number of invasive studies were included as a reference, but these are not part of the core review scope of this table. Metrics(Perceptual Similarity Metrics (PSM) include CLIP, C3D (Convolutional 3D), AlexNet, Inception, EffNet-B, SwAV). WER is Word Error Rate. BLEU is Bilingual Evaluation Understudy. METEOR is Metric for Evaluation of Translation with Explicit ORdering. ROUGE is Recall-Oriented Understudy for Gisting Evaluation. CIDEr is Consensus-based Image Description Evaluation. SPICE is Semantic Propositional Image Caption Evaluation. MFCC is Mel Frequency Cepstrum Coefficients. MCD is Mel-cepstral distortion. PCC is Pearson Correlation Coefficient. HER is Hand Error Rate. CER is Character-error-rate. STOI is ShortTime Objective Intelligibility.

Brain Modality	Decoded Content	Dataset	Representation Model	Generation Model	Work (Reference)	Method	Metrics
fMRI	Language	NSD (Allen et al. 2022)	CLIP	GPT-2	MindGPT (Chen et al. 2025b)	FMA	BLEU, ROUGE, METEOR, CIDEr, SPICE
			CLIP	GPT-2	Mind-SA (Chen et al. 2025a)	FMA	BLEU, ROUGE, METEOR, CIDEr, SPICE
			CoCa	-	BrainChat (Huang 2024)	FMA	PSM, METEOR, ROUGE, BLEU, CIDEr
			-	Vec2Text	From Eye to Mind (Feng and Nie 2025)	E2E	BLEU, METEOR, ROUGE, CIDEr, SPICE
			CLIP	LLaMA-3	NeuroCreat (Jing et al. 2025a)	FMA	PSM, PixCorr, SSIM, BLEU, ROUGE, METEOR, CIDEr, SPICE
			CoCa	Transformer	BrainChat (Huang et al. 2025)	FMA	PSM, METEOR, ROUGE
		Narratives fMRI (Nastase et al. 2021)	-	Transformer (Decoder)	PredFT (Yin et al. 2024)	E2E	BLEU, ROUGE
		CLSR (Tang et al. 2023)	-	GPT-1	(Tang et al. 2023)	AG-EML	WER, BLEU, METEOR, BERTScore, Human Acc.
		NL fMRI (LeBel et al. 2023)	-	GPT-2	BP-GPT (Chen et al. 2024b)	AG-EML	BLEU, METEOR, BERTScore
		COCO (Lin et al. 2014)	DINOv2	GPT-2	(Shen et al. 2025)	FMA	PSM, METEOR, ROUGE, Sentence
		HAD (Zhou et al. 2023)	SC-MBM	Video-LLaMA	LLM4Brain (Zheng and Sun 2024)	FMA	BERTScore, BLEU, ROUGE
	Speech	VGGSound (Chen et al. 2020)	VGGish	Audio Transformer	(Park et al. 2023)	DPC	Pearson Correlation, Pairwise Identification Acc.

Continued on next page...

Table 5 – Continued from previous page

Brain Modality	Decoded Content	Dataset	Representation Model	Generation Model	Work (Reference)	Method	Metrics
fMRI	Speech	Self-collected Dataset (de Zuazo et al. 2024)	Wav2vec2	Transformer	neural2speech (de Zuazo et al. 2024)	AG-EML	Acc.
EEG	Language	ZuCo (Hollenstein et al. 2018)	Convolutional Transformer	Multi-View Trans-former	EEG2TEXT (Liu et al. 2024a)	E2E	BLEU, ROUGE
			Multi-stream Transformer	-	CET-MAE (Wang et al. 2024)	FMA	BLEU, ROUGE
			CLIP	GIT	PinVC (Jing et al. 2025b)	FMA	BLEU, ROUGE
			VQ-VAE	BART	DeWave (Duan et al. 2023)	E2E	BLEU, ROUGE
		Self-collected Dataset (Lévy et al. 2025)	CNN + Transformer	-	Brain2Qwerty (Lévy et al. 2025)	DPC	HER, CER
	Speech	ISD (Xiong et al. 2025)	XI Model	XII Model	ETS (Xiong et al. 2025)	E2E	Spectrogram Similarity, MFCC Similarity, Human Acc.
		Self-collected Dataset (Abdulghani et al. 2023)	Wavelet Scattering Transformation	LSTM-RNN	(Abdulghani et al. 2023)	E2E	Acc., Precision, Recall, F-score
		Self-collected Dataset (Guo et al. 2023)	-	DualGAN	Dual-DualGAN (Guo et al. 2023)	E2E	Acc., MCD, PCC
		MOUS (Schoffelen et al. 2019)	Wav2vec2	-	(Défossez et al. 2023)	DPC	Identification Acc.
MEG	Speech	MEG Narrative (Armeni et al. 2022)	GAT	3D CNN	(Gideoni et al. 2024)	FMA	Balanced Acc.
		MOUS (Schoffelen et al. 2019)	Wav2vec2	-	(Défossez et al. 2023)	DPC	Identification Acc.
		Self-collected Dataset (de Zuazo et al. 2024)	Wav2vec2	Transformer	neural2speech (de Zuazo et al. 2024)	AG-EML	Acc.
	Language	Self-collected Dataset (Lévy et al. 2025)	CNN + Transformer	-	Brain2Qwerty (Lévy et al. 2025)	DPC	HER, CER
sEEG	Language	Self-collected Dataset (Pescatore et al. 2025)	Word2vec	-	(Pescatore et al. 2025)	FMA	Decoding Acc.

Continued on next page...

Table 5 – Continued from previous page

Brain Modality	Decoded Content	Dataset	Representation Model	Generation Model	Work (Reference)	Method	Metrics
sEEG	Speech	Self-collected Dataset (Chen et al. 2024a)	Swin Transformer	-	SwinTW (Chen et al. 2024a)	DPC	PCC, STOI, MCD
Intracranial EEG	Language	Self-collected Dataset (Willett et al. 2023)	RNN	-	(Willett et al. 2023)	E2E	Phoneme Error Rate, WER
ECoG	Speech	Self-collected Dataset (Chen et al. 2024a)	Swin Transformer	-	SwinTW (Chen et al. 2024a)	DPC	PCC, STOI, MCD
HD-DOT	Speech	HD-DOT(Tripathy et al. 2021)	Template Matching	-	(Tripathy et al. 2025)	FMA	Decoding Acc.

Reliability & reproducibility signals. Reliability is higher when multi-subject and cross-session validation is reported together with fully specified preprocessing and decoding hyperparameters. However, datasets and reporting practices remain heterogeneous (as reflected in Table 5), and code availability is uneven. We recommend documenting subject/session counts, split strategies, and paired reporting of literal fidelity and semantic consistency metrics.

5.4 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 (Park et al. 2023). 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 (Geirnaert et al. 2021). Early linear pipelines (stimulus/temporal response functions, CCA) (De Taillez et al. 2020) provided a principled starting point (Wong et al. 2018; Fiedler et al. 2017), but they struggle with nonlinearity and short decision windows that are critical for real-time control (Puffay et al. 2023; Cai et al. 2020; Vandecappelle et al. 2021). Recent systems replace these with non-linear encoders that borrow FM-inspired components—self-attention, convolutional hierarchies (Fan et al. 2024; Su et al. 2021; Xu et al. 2022; Accou et al. 2023), and graph neural networks—to model spatio-temporal structure and inter-sensor topology, yielding markedly higher accuracy at short latencies (Cai et al. 2023; Fan et al. 2025b,a; Li et al. 2025b). Within this family, some approaches perform identity-based AAD by correlating EEG with reference speech streams using CNN/attention encoders or encoder–decoders (Su et al. 2021; Xu et al. 2022; Accou et al. 2023); 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 (Accou et al. 2023; Fan et al. 2024; Zhou et al. 2025; Ni et al. 2024; Fan et al. 2025b; Yan et al. 2024; Fan

et al. 2025a). Collectively, these non-linear models narrow the gap to practical neuro-steered hearing aids by improving robustness in short windows (Geirnaert et al. 2021; Zhang et al. 2023).

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 (Park et al. 2023). 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 (Postolache et al. 2025; Simon et al. 2022). Notably, DMF2Mel (Fan et al. 2025c) 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 cortical processing pathways and with the added structural complexity of music (Simon et al. 2022). This complexity is being used as a probe for auditory scene analysis, while fMRI studies map genre-level representations across cortex (Nakai et al. 2021). 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. However, evidence under realistic acoustic mixtures, strict latency constraints, and cross-session robustness is still comparatively limited.

Failure modes & boundary conditions. AAD is constrained by decision-window length, acoustic scene complexity, and session variability; gains under controlled mixtures may not persist under realistic noise and artifacts. For auditory reconstruction, strong generative priors can mask weak neural evidence, and capturing long-range musical structure remains difficult from noisy measurements. In hearing-aid-like deployments, latency and calibration stability are often the bottlenecks.

Contradictions across studies. Comparisons differ due to stimulus type (speech and music), availability of clean references, window definitions, and metric choices (accuracy and correlation-based scores). Deep/FM-inspired models can show large gains in some controlled settings, while classical TRF/CCA baselines remain competitive when data are limited or interpretability and stable calibration matter. This motivates matched baselines and standardized protocols.

Reliability & reproducibility signals. Reliability improves when studies report cross-session validation, realistic mixtures, and strong matched baselines under consistent preprocessing. Reporting of attention labels, windowing, and latency/compute

constraints is still inconsistent, limiting reproducibility. We recommend standardizing window lengths, reporting latency, and including both classical and FM-based baselines.

5.5 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 (Xia et al. 2024). Complementary work explicitly aligns EEG, MEG, and fMRI representations in a common space to support decoding, encoding, and modality conversion, showing that an FM-style alignment layer can bridge disparate temporal/spatial scales and improve cross-subject generalization (Ferrante et al. 2024c). 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 (McNabb et al. 2025; Wang et al. 2022b). 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 (Li et al. 2025c).

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 (Fu et al. 2025b). 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 (Frey et al. 2021). 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.

5.6 Interpreting Metrics across Research

Tables 4 and 5 summarize the metrics reported in recent visual and language/speech decoding research. However, a key reason why quantitative synthesis remains difficult is that these metrics correspond to different evaluation objectives and are often calculated under incompatible protocols; therefore, the same metric names do not necessarily allow for comparison and support of the same conclusions.

We categorize the metrics into three types. The first is **fidelity**, which examines whether the decoded output is supported by neural evidence or matches explicit ground truth (e.g., SSIM / PSNR, PixCorr / PCC, N-way Accuracy for images; WER / CER/HER for speech under specific decoding constraints). The second is **perceptual quality** (e.g., FID / IS, Human Accuracy), which reflects realism or human subjective judgment or preference. However, it may be influenced by generative priors and therefore cannot serve as independent evidence for decoding evaluation. The third is **semantic consistency** (e.g., PSM / ROUGE / BLEU / METEOR / CIDEr / SPICE or BERTScore), which assesses meaning preservation. These three types of metrics can complement each other but are not interchangeable; even if fidelity decreases, perceptual metrics may improve. Strict pixel or n-gram overlap scores may penalize semantically correct outputs in cases of paraphrasing or appearance changes.

Incompatibility between different research findings also stems from data acquisition methods, candidate set construction for identification, preprocessing and channel count, feature extraction and normalization, and cueing and decoding settings. Therefore, interpretation should only be based on the reported protocol. We recommend that research reports validate the protocol, along with at least one fidelity-oriented metric and one semantic or utility-oriented metric. Key protocol details required for reproducible scoring should also be disclosed (e.g., grouping design, candidate set, and decoding/generation settings). Cross-paper comparisons should be limited to studies employing the same protocol and evaluating the same target attribute. Quantitative ranking between papers is only meaningful within a subset that employs the same protocol and consistently evaluates the target attribute.

6 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 (Gong et al. 2024; Sato and Kobayashi 2025; Park et al. 2023; Jayalath et al. 2024). 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 (Li et al. 2024; Chen et al. 2024b), (ii) direct regression that predicts FM embeddings or conditioning variables from neural signals, 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 (Tang et al. 2023).

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 (Tang et al. 2023; Gong et al. 2024). 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 (Li et al. 2024; Jayalath et al. 2024). 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 (Qiu et al. 2025a; Jayalath et al. 2024; Li et al. 2024; Tang et al. 2023).

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) (Qiu et al. 2025a; Xia et al. 2024). 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 (Benchetrit et al. 2023; Li et al. 2024; Jayalath et al. 2024). 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 (Gong et al. 2024; Li et al. 2024; Chen et al. 2023b).

7 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.

7.1 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 (Chen et al. 2023a; Jiang et al. 2024; Shirakawa et al. 2024). These constraints also shape the strength of evidence available in current FM-based decoding studies. We therefore distinguish four validation regimes with increasing translational relevance: within-session, single-subject evaluation; cross-session evaluation on the same subject (to probe drift); cross-subject evaluation (to assess population-level generalization); and naturalistic or online settings with latency and robustness constraints. Much of the literature remains concentrated in the first two regimes, whereas reproducible gains in cross-subject and deployment-oriented settings are less consistently demonstrated. This distinction is useful for interpreting "generalization" claims throughout this section.

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 (da Silva 2013; Logothetis 2008).

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 (i.e., Level-3 cross-subject validation) 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 (Luo et al. 2024; Chen et al. 2023a).

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 which becomes most stringent under Level-4 online or naturalistic constraints. 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 (Gong et al. 2024).

Consequently, the black-box opaqueness of these models exposes a vulnerability to targeted adversarial attacks. Many recent systems rely on an alignment step which

maps brain signals into a high-dimensional latent space, used for retrieval or conditional generation. This acts as an attack surface where adversarial perturbations can be crafted to induce small yet semantically meaningful shifts in the representation, altering retrieval outcomes or the conditioning signals passed to a generator (Zhang et al. 2020; Lubars and Chin 2022; Li et al. 2021). As generative models are able to produce fluent text and images given weak or ambiguous input, this induced misalignment often manifests as a high-confidence but semantically incorrect - hallucinated - output rather than an obvious corruption (Ji et al. 2023). As discussed earlier, risks of both test-time evasions and training-time backdoors are established threats in the literature (Meng et al. 2022). Therefore, robustness claims for FM-based decoders must explicitly state the assumed threat model and report rigorous adversarial stress-testing. Future work to quantify attack success rates under evasion and adversarial perturbations, alongside basic backdoor checks, would be recommended to enhance standard denoising pipelines to account for intentional, alongside accidental, noise.

At the same time, reproducibility and reporting are considerable bottlenecks in non-invasive brain decoding. In non-invasive brain decoding, claims are only as reproducible as the specificity with which the full experimental and computational pipeline is documented. Authors should therefore make the data provenance and cohort structure explicit (dataset release or acquisition protocol, access/consent constraints, subject and session counts, inclusion/exclusion rules, and a precise task specification with timing and label definitions), and provide a technically complete description of acquisition and preprocessing (instrumentation, sampling rate/TR, channel/voxel specification and referencing, followed by the exact sequence of denoising, artifact handling, normalization/alignment, and any modality-specific steps such as fMRI motion correction, together with ROI/voxel/channel selection and time-windowing). Equally important is a split protocol described at a level that rules out leakage (within-subject, cross-session, or cross-subject, with fold construction stated), since this choice often dominates reported performance; these split choices correspond directly to the validation regimes above and should be treated as part of the primary evidence claim. Method reporting should not stop at model names: the FM backbone version, alignment target space, optimization setup, and random seeds must be specified; when a generative component is used, prompts and decoding settings should be treated as part of the method and reported accordingly. Finally, evaluation should define metrics in relation to the quantity of interest (fidelity versus perceptual quality versus semantic consistency) and, where feasible, provide uncertainty estimates; releasing code, preprocessing scripts, configurations, and checkpoints remains the most direct mechanism for enabling independent verification.

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. Recent neuro-rights proposals argue that neural data processing implicates rights such as mental privacy and cognitive liberty (Ienca and Andorno 2017; Yuste et al. 2017), motivating protections beyond conventional biomedical data governance (European Parliament 2024;

Table 6. Operating regimes and validation recommendations for FM-based and classical BCI pipelines. Acc: accuracy; ITR: information transfer rate; Corr: stimulus–response correlation; Fidelity metrics include SSIM/PSNR (vision) or WER/CER (speech); Semantic metrics include CLIPScore/BERTScore or retrieval-based consistency.

Deployment Regime	Classical Baseline Strengths	FM-based Advantages	FM Failure Mode	Recommended Validation
Low-latency closed-set control (P300/SSVEP/MI)	Stable Calibration; Low Compute	Marginal Benefit	Prompt/Decoding Sensitivity; Calibration Overhead	Cross-session / Acc·ITR / Latency
Small-data; High-noise; Online Drift	Sample-efficient; Robust Regularization	Priors help if aligned	Alignment Drift; Domain Shift	Cross-session / AUROC·Corr / Drift stress-test
Clinical Interpretability / Auditability	Transparent Assumptions; Inspectable Parameters	Semantic Scaffolding	Plausible but Unfaithful Generation	Cross-session / Fidelity+Semantic / Blinded human eval
Open-vocabulary Semantic Decoding (Naturalistic)	Limited Open-set Semantics	Strong Semantic Priors; Retrieval	Hallucination; Metric Mismatch	Cross-subject / Semantic+Fidelity / Prompt sensitivity

UNESCO 2023). This challenges the adequacy of a static consent system where a single, one-off signature is obtained. As alignment techniques, prompting interfaces and downstream reuse evolve, decoders may enable secondary inferences beyond the user's original intent (F. et al. 2019; Future of Privacy Forum and IBM 2021) such as biometric 'brainprints' (Yang et al. 2022), mood/intent inference, or other latent attributes. A practical path forward is adoption of a tiered consent system which explicitly separates permissible classes of outputs/inferences (e.g., task labels vs. semantic reconstructions vs. affect/clinical inference), alongside dynamic consent mechanisms which support ongoing re-consent, granular preference updates, withdrawal and audit for when the use cases change over time (Kaye et al. 2015; Budin-Ljøsne et al. 2017; Wiertz and Boldt 2022).

7.2 Applicability Boundaries for FM-Based BCI

Foundation Model brain decoding is usually built around two pieces: a learned mapping from brain activity into an intermediate representational space, and a strong generative prior that turns that representation into a perceptible output. This design can noticeably improve output richness in controlled experiments, but its reliability is uneven once you move across subjects, across sessions, or into settings with tight deployment constraints. Performance often drops when the alignment itself drifts (e.g., day-to-day or subject-to-subject variability), when the generator "fills in" missing evidence with plausible content, or when results hinge on prompting choices and decoding hyperparameters. Taken together, these limitations argue for stating the operating boundaries explicitly, rather than treating FM-based pipelines as a universal upgrade.

Classical pipelines remain the better default in several high-value regimes. For low-latency, closed-set control tasks (P300/SSVEP/MI), lightweight classifiers paired with well-understood signal-processing steps are typically more stable and easier

Table 7. Prioritized research roadmap for foundation-model–based non-invasive brain decoding.

Priority	Research Objective	Methodology	Reference Benchmarks
P1	Benchmarking and comparability	Fixed splits; Matched baselines; Open evaluation scripts	NSD, THINGS, Narratives, Things-EEG2
P2	Cross-session and cross-subject robustness	Functional alignment; PEFT/adapters; Domain adaptation; Drift-robust training	Narratives, MOUS, HCP-style
P3	Real-time and resource-constrained deployment	Streaming encoders; Distillation/quantization; Memory-efficient generation	AAD corpora, MOABB paradigms, portable EEG tasks
P4	Trustworthy decoding and governance	Threat modeling; Red-teaming; Privacy-preserving learning	Any benchmark + Standardized misuse tests

to deploy than generative decoding. When data are scarce and drift is unavoidable, sample-efficient and interpretable baselines, such as TRF/mTRF and CCA in auditory attention decoding—are often easier to regularize and to validate under cross-session evaluation. In MI classification, covariance-based Riemannian methods have a long track record of robust performance with modest data requirements (See Table 6 for details). Across these cases, the priority is reliability, calibration stability, and latency; adding large generative components can increase complexity without delivering consistent gains.

The same ingredients that make FM-based pipelines attractive can also make failure modes sharper. If the brain-to-semantic mapping is even slightly misaligned, the downstream generator may amplify that error into a confident-looking output. Moreover, a strong prior can mask weak neural evidence, producing results that score well on perceptual criteria yet deviate from the underlying signal—especially when evaluation emphasizes plausibility over neural fidelity. Finally, both LLM- and diffusion-based modules are sensitive to prompting and decoding settings, which can introduce avoidable variance unless standardized. For these reasons, FM-based improvements should be demonstrated under matched latency budgets, explicit cross-session and cross-subject splits, and a metric suite that jointly captures fidelity and semantic consistency.

7.3 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 pre-trained 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 (Jayalath et al. 2024), 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 (Huth et al. 2016). 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 (Tang et al. 2023).

This roadmap is based on the premise that evaluation results in this field are comparable across different studies and meaningful for real-world deployment. Currently, many studies obtain results under different experimental protocols, such as different subjects, preprocessing procedures, and decoding constraints. Furthermore, the metrics used also vary, thus failing to reflect the same underlying conclusions (Jayaram and Barachant 2018). This is crucial for practical applications in brain decoding; only through explicit cross-session and cross-subject validation can robustness to drift, low-latency performance, and calibration burden be reliably assessed (Lotte et al. 2007). Ideally, stress tests simulating natural conditions should also be conducted. Therefore, we recommend and advocate for the use of a minimal benchmarking protocol for future FM-based BCI research: 1) report at least one cross-session evaluation and, where feasible, one cross-subject evaluation; 2) compare against a highly matched baseline under the same preprocessing conditions (Glaser et al. 2019); 3) document the cueing and decoding/generation settings as part of the methodology; and 4) use a set of metrics that distinguishes between fidelity, semantic consistency, and task usefulness, while reporting uncertainty. In addition to making different studies comparable, this standardized benchmarking scheme also lays the foundation for subsequent discussions on deployment and dual-use, making empirical capabilities more clearly defined. For clarity, Table 7 provides a prioritized roadmap connecting major obstacles with research objective, actionable methodology and reference benchmarks. This roadmap emphasizes that comparability (P1) is a prerequisite for assessing robustness (P2) and deployment feasibility (P3), while also regarding credible decoding and governance (P4) as equally important objectives, rather than as ex-post considerations. These directions collectively put the evidence hierarchy discussed above into practice, clarifying the assessment mechanisms and reporting expectations.

Against this backdrop, while neural decoding holds clear clinical promise such as in communication interfaces (Anumanchipalli et al. 2019; Willett et al. 2021a) and rehabilitation in severe motor impairment (Chaudhary et al. 2016), the same representational power can be repurposed for non-therapeutic and potentially coercive uses. This creates a dual-use dilemma where advances that improve decoding fidelity, generalisation and robustness also increase the plausibility of surveillance applications (Zuboff 2019). This includes attempts to infer sensitive mental content from non-consenting subjects, as well as workplace and consumer monitoring (Future of Privacy Forum and IBM 2021) and 'neuromarketing'-style behavioural optimisation. These capabilities in strategic contexts raise questions about military and intelligence adoption pathways even when current systems remain technically constrained (Claverie and du Cluzel 2022; Defense Advanced Research Projects Agency (DARPA) 2018). From a research perspective, this motivates treating misuse-resistance as a priority

objective alongside accuracy. Concrete directions include explicit threat modelling and red-teaming for coercive or non-consensual settings, governance and deployment constraints such as purpose limitation, access control and secure logging (European Parliament 2024; UNESCO 2023). Finally, the use of technical safeguards such as encryption and differential privacy which can reduce unintended inference and unauthorized reuse across diverse tasks or contexts (Bonaci et al. 2014; Future of Privacy Forum and IBM 2021).

8 Conclusion

Foundation models have substantially accelerated progress in non-invasive brain decoding by providing transferable representations and strong generative priors that improve semantic alignment and reconstruction quality across modalities. 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. Collectively, these advances broaden what is achievable in controlled settings, while robust cross-subject generalization and deployment-level reliability remain open challenges.

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.

In summary, FM-enabled decoding represents a compelling path toward assistive communication (Anumanchipalli et al. 2019; Willett et al. 2021a) and rehabilitation (Chaudhary et al. 2016), but its societal impact will depend on how capability is governed and constrained. As models become more general and decoding outputs more semantically expressive, the boundary between clinical assistive use and surveillance-oriented inference can narrow. This risk is particularly acute when neural data is linked to identity and used beyond the original task's context (Future of Privacy Forum and

IBM 2021), highlighting the need for robust institutional oversight (European Parliament 2024; UNESCO 2023). A mature research agenda should therefore explicitly acknowledge the dual-use potential and incorporate safety-by-design principles accordingly. Technical progress must be aligned with protections for mental privacy and cognitive liberty (Yuste et al. 2017; European Parliament 2024; UNESCO 2023), ensuring that sufficient technical safeguards are embedded with auditable accountability mechanisms from the development phase (Bonaci et al. 2014).

Acknowledgements. 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.

Declarations

- **Funding:** 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).
- **Conflict of interest:** The authors have no relevant financial or non-financial interests to disclose.
- **Ethics approval and consent to participate:** Not applicable. This article is a review/survey study and does not involve new studies with human participants or animals performed by any of the authors.
- **Consent for publication:** Not applicable.
- **Data availability:** No new data were generated or analysed in this study. This article is a survey of previously published literature, and dataset information summarized in Table 2 is derived from the cited sources.
- **Materials availability:** Not applicable.
- **Code availability:** No code was generated or used in this study.
- **Author contribution:** 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. Wenhao Cai: Visualization. George Ford: Writing – review and editing. Yang Cui: Writing – review and editing. Yunhao Zhang: Methodology, 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.

References

Abdulghani MM, Walters WL, Abed KH. Imagined speech classification using EEG and deep learning. Bioengineering. 2023;10(6):649. https://www.mdpi.com/2306-5354/10/6/649.

Abiri R, Borhani S, Sellers EW, Jiang Y, Zhao X. A comprehensive review of EEG-based brain–computer interface paradigms. Journal of neural engineering. 2019;16(1):011001. https://iopscience.iop.org/article/10.1088/1741-2552/aaf12e/meta.

Accou B, Vanthornhout J, hamme HV, Francart T. Decoding of the speech envelope from EEG using the VLAAI deep neural network. Scientific Reports. 2023;13(1):812. https://www.nature.com/articles/s41598-022-27332-2.

Allen EJ, St-Yves G, Wu Y, Breedlove JL, Prince JS, Dowdle LT, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience. 2022;25(1):116–126. https://www.nature.com/articles/s41593-021-00962-x.

Altaheri H, Muhammad G, Alsulaiman M, Amin SU, Altuwaijri GA, Abdul W, et al. Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review. Neural Computing and Applications. 2023;35(20):14681–14722. https://link.springer.com/article/10.1007/s00521-021-06352-5.

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

Anumanchipalli GK, Chartier J, Chang EF. Speech synthesis from neural decoding of spoken sentences. Nature. 2019;568:493–498.

Armeni K, Güçlü U, van Gerven M, Schoffelen JM. A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. Scientific Data. 2022;9(1):278. https://www.nature.com/articles/s41597-022-01382-7.

Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing

systems. 2020;33:12449–12460.

Bai Y, Wang X, Cao Yp, Ge Y, Yuan C, Shan Y. Dreamdiffusion: Generating high-quality images from brain eeg signals. arXiv preprint arXiv:230616934. 2023;https://arxiv.org/abs/2306.16934.

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

Benchetrit Y, Banville H, King JR. Brain decoding: toward real-time reconstruction of visual perception. arXiv preprint arXiv:231019812. 2023;https://arxiv.org/abs/2310.19812.

Bonaci T, Calo R, Chizeck HJ. App Stores for the Brain: Privacy & Security in Brain-Computer Interfaces. In: Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM Workshops) Budapest, Hungary; 2014. p. 510–514.

Bottemanne H. Bayesian brain theory: Computational neuroscience of belief. Neuroscience. 2025;566:198–204.

Budin-Ljøsne I, Teare HJA, Kaye J, Beck S, Bentzen HB, Caenazzo L, et al. Dynamic consent: a potential solution to some of the challenges of modern biomedical research. BMC Medical Ethics. 2017;18(1):4. https://doi.org/10.1186/s12910-016-0162-9, https://doi.org/10.1186/s12910-016-0162-9.

Buzsáki G, Anastassiou CA, Koch C. The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. Nature reviews neuroscience. 2012;13(6):407–420. https://www.nature.com/articles/nrn3241.

Cai S, Schultz T, Li H. Brain topology modeling with EEG-graphs for auditory spatial attention detection. IEEE Transactions on Biomedical Engineering. 2023;71(1):171–182. https://ieeexplore.ieee.org/abstract/document/10178086/.

Cai S, Su E, Song Y, Xie L, Li H. Low Latency Auditory Attention Detection with Common Spatial Pattern Analysis of EEG Signals. In: INTERSPEECH; 2020. p. 2772–2776. http://www.interspeech2020.org/uploadfile/pdf/Wed-SS-2-7-3.pdf.

Cao X, Gong P, Zhang L, Zhang D. EEG-CLIP: A transformer-based framework for EEG-guided image generation. Neural Networks. 2025;p. 108167.

Carìa A. 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; 2025. p. 403–416. https://link.springer.com/chapter/10.1007/978-3-031-84457-7_25.

Chaudhary U, Birbaumer N, Ramos-Murguialday A. Brain-computer interfaces for communication and rehabilitation. Nat Rev Neurol. 2016;12:513–525.

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

Chen J, Qi Y, Wang Y, Pan G. 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; 2025. p. 21938–21948.

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

Chen J, Chen X, Wang R, Le C, Khalilian-Gourtani A, Jensen E, et al. Subject-agnostic transformer-based neural speech decoding from surface and depth electrode signals. bioRxiv. 2024;https://pmc.ncbi.nlm.nih.gov/articles/PMC10980022/.

Chen X, Du C, Liu C, Wang Y, He H. Open-vocabulary Auditory Neural Decoding Using fMRI-prompted LLM. arXiv preprint arXiv:240507840. 2024;https://arxiv.org/abs/2405.07840.

Chen Z, Qing J, Xiang T, Yue WL, Zhou JH. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. 2023;p. 22710–22720. https://arxiv.org/abs/2211.06956.

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

Claverie B, du Cluzel F.: The Cognitive Warfare Concept; 2022. https://www.innovationhub-act.org/sites/default/files/2022-02/CW%20article%20Claverie%20du%20Cluzel%20final_0.pdf, accessed: 2025. NATO Allied Command Transformation Innovation Hub.

Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: a review. Journal of neural engineering. 2019;16(3):031001. https://iopscience.iop.org/article/10.1088/1741-2552/ab0ab5/meta.

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

De Taillez T, Kollmeier B, Meyer BT. Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. European Journal of Neuroscience. 2020;51(5):1234–1241. https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.13790.

Defense Advanced Research Projects Agency (DARPA).: Nonsurgical Neural Interfaces Could Significantly Expand Use of Neurotechnology; 2018. https://www.darpa.mil/news/2018/nonsurgical-neural-interfaces, accessed: 2026-02-15.

Défossez A, Caucheteux C, Rapin J, Kabeli O, King JR. Decoding speech perception from non-invasive brain recordings. Nature Machine Intelligence. 2023;5(10):1097–1107. https://www.nature.com/articles/s42256-023-00714-5.

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

Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019;p. 4171–4186. https://arxiv.org/abs/1810.04805.

Dmochowski JP, Ki JJ, DeGuzman P, Sajda P, Parra LC. Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity. NeuroImage. 2018;180:134–146. https://www.sciencedirect.com/science/article/pii/S1053811917304299.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020;.

Du B, Cheng X, Duan Y, Ning H. fmri brain decoding and its applications in brain–computer interface: A survey. Brain Sciences. 2022;12(2):228.

Duan Y, Zhou J, Wang Z, Wang YK, Lin CT. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. arXiv preprint arXiv:230914030. 2023;https://arxiv.org/abs/2309.14030.

Eickenberg M, Gramfort A, Varoquaux G, Thirion B. Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage. 2017;152:184–194.

Elam JS, Glasser MF, Harms MP, Sotiropoulos SN, Andersson JL, Burgess GC, et al. The human connectome project: a retrospective. NeuroImage. 2021;244:118543. https://www.sciencedirect.com/science/article/pii/S1053811921008168.

European Parliament. The protection of mental privacy in the area of neuroscience. Brussels, Belgium: Panel for the Future of Science and Technology (STOA); 2024.

F C, V G, G G, I D. Brain-Data Privacy: A Study on the Risks of Inference on Shared BCI Datasets. In: Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) Berlin, Germany; 2019. p. 4471–4474.

Fan C, Yang X, Zhang H, Chen Y, Li L, Zhou J, et al. 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; 2025. p. 4137–4145. https://doi.org/10.24963/ijcai.2025/461.

Fan C, Zhang H, Huang W, Xue J, Tao J, Yi J, et al. DGSD: Dynamical graph self-distillation for EEG-based auditory spatial attention detection. Neural Networks. 2024;179:106580. https://www.sciencedirect.com/science/article/abs/pii/S0893608024005045.

Fan C, Zhang H, Ni Q, Zhang J, Tao J, Zhou J, et al. Seeing helps hearing: A multi-modal dataset and a mamba-based dual branch parallel network for auditory attention decoding. Information Fusion. 2025;p. 102946. https://www.sciencedirect.com/science/article/abs/pii/S1566253525000193.

Fan C, Zhang S, Zhang J, Liu E, Li X, Zhao G, et al. DMF2Mel: A Dynamic Multiscale Fusion Network for EEG-Driven Mel Spectrogram Reconstruction. In: Proceedings of the 33rd ACM International Conference on Multimedia; 2025. p. 6977–6985. https://doi.org/10.1145/3746027.3755501.

Fei T, Uppal A, Jackson IL, Wang D, de Sa VR. Perceptogram: Visual Reconstruction from EEG Using Image Generative Models. 2024;https://openreview.net/forum?id=IZOeRDS6zU.

Feng F, Nie J. From Eye to Mind: brain2text Decoding Reveals the Neural Mechanisms of Visual Semantic Processing. arXiv preprint arXiv:250322697. 2025;https://arxiv.org/abs/2503.22697.

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

Ferrante M, Boccato T, Passamonti L, Toschi N. Retrieving and reconstructing conceptually similar images from fMRI with latent diffusion models and a neuro-inspired brain decoding model. Journal of Neural Engineering. 2024;21(4):046001. https://iopscience.iop.org/article/10.1088/1741-2552/ad593c/meta.

Ferrante M, Boccato T, Rashkov G, Toschi N. Towards Neural Foundation Models for Vision: Aligning EEG, MEG, and fMRI Representations for Decoding, Encoding, and Modality Conversion. arXiv preprint arXiv:241109723. 2024;https://arxiv.org/

abs/2411.09723.

Ferrante M, Boccato T, Toschi N. Semantic brain decoding: from fmri to conceptually similar image reconstruction of visual stimuli. arXiv preprint arXiv:221206726. 2022;https://arxiv.org/abs/2212.06726.

Ferrante M, Ozcelik F, Boccato T, VanRullen R, Toschi N. Brain captioning: Decoding human brain activity into images and text. arXiv preprint arXiv:230511560. 2023;https://arxiv.org/abs/2305.11560.

Ferrari M, Quaresima V. A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. Neuroimage. 2012;63(2):921–935. https://www.sciencedirect.com/science/article/abs/pii/S1053811912003308.

Fiedler L, Wöstmann M, Graversen C, Brandmeyer A, Lunner T, Obleser J. Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. Journal of neural engineering. 2017;14(3):036020. https://iopscience.iop.org/article/10.1088/1741-2552/aa66dd/meta.

Frey M, Nau M, Doeller CF. Magnetic resonance-based eye tracking using deep neural networks. Nature neuroscience. 2021;24(12):1772–1779.

Friston K. The free-energy principle: a rough guide to the brain? Trends in cognitive sciences. 2009;13(7):293–301.

Fu H, Wang H, Chin JJ, Shen Z. 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; 2025. p. 1–5.

Fu M, Chen G, Zhang Y, Zhang M, Wang Y. Comprehensive Neural Representations of Naturalistic Stimuli through Multimodal Deep Learning. bioRxiv. 2025;p. 2025–04.

Future of Privacy Forum, IBM. Privacy and Ethical Recommendations for Brain-Computer Interfaces. Washington, DC, USA: Future of Privacy Forum; 2021.

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

Gao J, Fu Y, Fu Y, Wang Y, Qian X, Feng J. 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. 2025;.

Geirnaert S, Vandecappelle S, Alickovic E, De Cheveigne A, Lalor E, Meyer BT, et al. Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices. IEEE Signal Processing Magazine. 2021;38(4):89–102. https://ieeexplore.ieee.org/abstract/document/9467380.

Gideoni Y, Timms RC, Jones OP. Non-invasive Neural Decoding in Source Reconstructed Brain Space. arXiv preprint arXiv:241019838. 2024;https://arxiv.org/abs/2410.19838.

Gifford AT, Dwivedi K, Roig G, Cichy RM. A large and rich EEG dataset for modeling human visual object recognition. NeuroImage. 2022;264:119754. https://www.sciencedirect.com/science/article/pii/S1053811922008758.

Glaser JI, Benjamin AS, Farhoodi R, Kording KP. The roles of supervised machine learning in systems neuroscience. Progress in neurobiology. 2019;175:126–137.

Gong Z, Bao G, Zhang Q, Wan Z, Miao D, Wang S, et al. NeuroClips: Towards high-fidelity and smooth fMRI-to-video reconstruction. Advances in Neural Information Processing Systems. 2024;37:51655–51683. https://proceedings.neurips.cc/paper_files/paper/2024/hash/5c594bf6223b67109441c9e0c97542ed-Abstract-Conference.html.

Gong Z, Zhang Q, Bao G, Zhu L, Xu R, Liu K, et al. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39; 2025. p. 14247–14255.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Communications of the ACM. 2020;63(11):139–144.

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

Han K, Wen H, Shi J, Lu KH, Zhang Y, Fu D, et al. Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. NeuroImage. 2019;198:125–136. https://www.sciencedirect.com/science/article/abs/pii/S1053811919304318.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 2001;293(5539):2425–2430.

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

2022_paper.

Hebart MN, Contier O, Teichmann L, Rockter AH, Zheng CY, Kidder A, et al. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. Elife. 2023;12:e82580. https://elifesciences.org/articles/82580.

Henighan T, Kaplan J, Katz M, Chen M, Hesse C, Jackson J, et al. Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:201014701. 2020;https://arxiv.org/abs/2010.14701.

Hestness J, Narang S, Ardalani N, Diamos G, Jun H, Kianinejad H, et al. Deep learning scaling is predictable. Empirically arXiv. 2017;1712:2. https://arxiv.org/abs/1712.00409.

Hollenstein N, Rotsztejn J, Troendle M, Pedroni A, Zhang C, Langer N. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. Scientific data. 2018;5(1):1–13. https://www.nature.com/articles/sdata2018291.

Horikawa T, Kamitani Y. Generic decoding of seen and imagined objects using hierarchical visual features. Nature communications. 2017;8(1):15037. https://www.nature.com/articles/ncomms15037.

Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural networks. 1989;2(5):359–366.

Huang W. BrainChat: Decoding Semantic Information from fMRI using Vision-language Pretrained Models. arXiv preprint arXiv:240607584. 2024;https://arxiv.org/abs/2406.07584.

Huang W, Ma K, Xie T, Wang H. 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; 2025. p. 1–5.

Huth AG, De Heer WA, Griffiths TL, Theunissen FE, Gallant JL. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature. 2016;532(7600):453–458. https://www.nature.com/articles/nature17637.

Ienca M, Andorno R. Towards new human rights in the age of neuroscience and neurotechnology. Life Sciences, Society and Policy. 2017;13(1):5. https://doi.org/10.1186/s40504-017-0050-1, https://doi.org/10.1186/s40504-017-0050-1.

Jayalath D, Landau G, Shillingford B, Woolrich M, Jones OP. The Brain's Bitter Lesson: Scaling Speech Decoding With Self-Supervised Learning. arXiv preprint arXiv:240604328. 2024;https://arxiv.org/abs/2406.04328.

Jayaram V, Barachant A. MOABB: trustworthy algorithm benchmarking for BCIs. Journal of neural engineering. 2018;15(6):066011.

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys. 2023 Mar;55(12). https://doi.org/10.1145/3571730.

Jiang S, Li H, Meng Z, Su F, Zhao Z. Mind2Word: Towards Generalized Visual Neural Representations for High-Quality Video Reconstruction. Expert Systems with Applications. 2025;p. 129154.

Jiang WB, Zhao LM, Lu BL. Large brain model for learning generic representations with tremendous EEG data in BCI. arXiv preprint arXiv:240518765. 2024;https://arxiv.org/abs/2405.18765.

Jiao Z, You H, Yang F, Li X, Zhang H, Shen D. Decoding EEG by Visual-guided Deep Neural Networks. In: IJCAI, vol. 28 Macao; 2019. p. 1387–1393. https://www.ijcai.org/proceedings/2019/0192.pdf.

Jing H, Jiang D, Ma Y, Hua H, Huang B, Zheng N. Beyond Brain Decoding: Visual-Semantic Reconstructions to Mental Creation Extension Based on fMRI. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2025. p. 19258–19268.

Jing H, Ma Y, Yang P, Hua H, Zheng N. Pinpointing Visual Content: Disentangled Features in Multimodal Model for EEG Representation Learning and Decoding. Knowledge-Based Systems. 2025;p. 114212.

Kalantari F, Faez K, Amindavar H, Nazari S. Improved image reconstruction from brain activity through automatic image captioning. Scientific Reports. 2025;15(1):4907. https://www.nature.com/articles/s41598-025-89242-3.

Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. arXiv preprint arXiv:200108361. 2020;https://arxiv.org/pdf/2001.08361/1000.

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

Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008;452(7185):352–355. https://www.nature.com/articles/nature06713.

Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. European Journal of

Human Genetics. 2015;23(2):141–146. https://doi.org/10.1038/ejhg.2014.71, https://doi.org/10.1038/ejhg.2014.71.

Khare S, Choubey RN, Amar L, Udutalapalli V. Neurovision: perceived image regeneration using cprogan. Neural Computing and Applications. 2022;34(8):5979–5991. https://link.springer.com/article/10.1007/s00521-021-06774-1.

Kim BH, Choi J, Yun E, Kim K, Li X, Lee J. Large-scale graph representation learning of dynamic brain connectome with transformers. arXiv preprint arXiv:231214939. 2023;.

Kim JM, Choi W, Park SJ, Heo KS, Son YH, Oh JH, et al. 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; 2025. p. 4824–4833.

Kriegeskorte N, Mur M, Bandettini PA. Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience. 2008;2:249.

Kupershmidt G, Beliy R, Gaziv G, Irani M. A penny for your (visual) thoughts: Self-supervised reconstruction of natural movies from brain activity. arXiv preprint arXiv:220603544. 2022;https://arxiv.org/abs/2206.03544.

Lan YT, Ren K, Wang Y, Zheng WL, Li D, Lu BL, et al. Seeing through the brain: image reconstruction of visual perception from human brain signals. arXiv preprint arXiv:230802510. 2023;https://arxiv.org/abs/2308.02510.

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

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

LeBel A, Wagner L, Jain S, Adhikari-Desai A, Gupta B, Morgenthal A, et al. A natural language fMRI dataset for voxelwise encoding models. Scientific Data. 2023;10(1):555. https://www.nature.com/articles/s41597-023-02437-z.

Lévy J, Zhang M, Pinet S, Rapin J, Banville H, d'Ascoli S, et al. Brain-to-text decoding: A non-invasive approach via typing. arXiv preprint arXiv:250217480. 2025;https://arxiv.org/abs/2502.17480.

Li D, Wei C, Li S, Zou J, Qin H, Liu Q. Visual decoding and reconstruction via eeg embeddings with guided diffusion. arXiv preprint arXiv:240307721. 2024;https://arxiv.org/abs/2403.07721.

Li H, Wu H, Chen B. NeuralDiffuser: Neuroscience-Inspired Diffusion Guidance for fMRI Visual Reconstruction. IEEE Transactions on Image Processing. 2025;.

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

Li L, Fan C, Zhang H, Zhang J, Yang X, Zhou J, et al. MHANet: Multi-scale Hybrid Attention Network for Auditory Attention Detection. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25; 2025. p. 4173–4181. https://doi.org/10.24963/ijcai.2025/465.

Li X, et al. Theory-based Vulnerability Analysis of Deep Learning-based Brain-Computer Interfaces. IEEE Trans Inf Forensics Secur. 2021;16:2378–2391.

Li Y, Zhu L, Huang A, Zhang J, Yuan P. Multimodal MBC-ATT: cross-modality attentional fusion of EEG-fNIRS for cognitive state decoding. Frontiers in Human Neuroscience. 2025;19:1660532.

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

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. 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; 2014. p. 740–755. https://arxiv.org/abs/1405.0312.

Liu H, Hajialigol D, Antony B, Han A, Wang X. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pre-training and multi-view transformer. arXiv preprint arXiv:240502165. 2024;https://arxiv.org/abs/2405.02165.

Liu P, Dong G, Guo D, Li K, Li F, Yang X, et al. A Survey on fMRI-based Brain Decoding for Reconstructing Multimodal Stimuli. arXiv preprint arXiv:250315978. 2025;https://arxiv.org/abs/2503.15978.

Liu XH, Liu YK, Wang Y, Ren K, Shi H, Wang Z, et al. EEG2video: Towards decoding dynamic visual perception from EEG signals. Advances in Neural Information Processing Systems. 2024;37:72245–72273. https://proceedings.neurips.cc/paper_files/paper/2024/hash/84bad835faaf48f24d990072bb5b80ee-Abstract-Conference.html.

Liu XH, Lu BL, Zheng WL. 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; 2025. p. 18273–18283.

Liu Y, Ma Y, Zhou W, Zhu G, Zheng N. BrainCLIP: Bridging brain and visual-linguistic representation via CLIP for generic natural visual stimulus decoding. arXiv preprint arXiv:230212971. 2023;https://arxiv.org/abs/2302.12971.

Liu Y, Ma Y, Zhu G, Jing H, Zheng N. See Through Their Minds: Learning Transferable Brain Decoding Models from Cross-Subject fMRI. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39; 2025. p. 5730–5738.

Logothetis NK. What we can do and what we cannot do with fMRI. Nature. 2008;453(7197):869–878. https://www.nature.com/articles/nature06976.

Lopez E, Sigillo L, Colonnese F, Panella M, Comminiello D. 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; 2025. p. 1–5.

Lotte F, Congedo M, Lécuyer A, Lamarche F, Arnaldi B. A review of classification algorithms for EEG-based brain–computer interfaces. Journal of neural engineering. 2007;4(2):R1–R13.

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

Lubars B, Chin TJ. Universal Adversarial Perturbations for EEG-Based Brain-Computer Interfaces. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Singapore; 2022. .

Luo Y, Liu W, Li H, Lu Y, Lu BL. A cross-scenario and cross-subject domain adaptation method for driving fatigue detection. Journal of Neural Engineering. 2024;21(4):046004. https://iopscience.iop.org/article/10.1088/1741-2552/ad546d/meta.

Mai W, Zhang Z. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. arXiv preprint arXiv:230807428. 2023;https://arxiv.org/abs/2308.07428.

McGeady C, Vučković A, Puthusserypady S. 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; 2019. p. 1–6. https://ieeexplore.ieee.org/abstract/document/8737333.

McNabb CB, Driver ID, Hyde V, Hughes G, Chandler HL, Thomas H, et al. WAND: A multi-modal dataset integrating advanced MRI, MEG, and TMS for multi-scale brain analysis. Scientific Data. 2025;12(1):220.

Mehmood T, Ahmad H, Shakeel MH, Taj M. CATVis: Context-Aware Thought Visualization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention Springer; 2025. p. 98–108.

Meng L, et al. EEG-Based Brain-Computer Interfaces Are Vulnerable to Backdoor Attacks. IEEE Trans Dependable Secur Comput. 2022;.

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

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, et al. Predicting human brain activity associated with the meanings of nouns. science. 2008;320(5880):1191–1195. https://www.science.org/doi/abs/10.1126/science.1152876.

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

Nakai T, Koide-Majima N, Nishimoto S. Correspondence of categorical and feature-based representations of music in the human brain. Brain and behavior. 2021;11(1):e01936. https://onlinelibrary.wiley.com/doi/full/10.1002/brb3.1936.

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL. Bayesian reconstruction of natural images from human brain activity. Neuron. 2009;63(6):902–915. https://www.cell.com/neuron/fulltext/S0896-6273(09)00685-0.

Nastase SA, Liu YF, Hillman H, Zadbood A, Hasenfratz L, Keshavarzian N, et al. The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. Scientific data. 2021;8(1):250. https://www.nature.com/articles/s41597-021-01033-3.

Ni Q, Zhang H, Fan C, Pei S, Zhou C, Lv Z. 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; 2024. p. 3115–3123. https://doi.org/10.24963/ijcai.2024/345.

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies. Current biology. 2011;21(19):1641–1646. https://www.sciencedirect.com/science/article/pii/S0960982211009377.

Olshausen BA, Field DJ. Sparse coding of sensory inputs. Current opinion in neurobiology. 2004;14(4):481–487.

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

Ozcelik F, VanRullen R. Natural scene reconstruction from fMRI signals using generative latent diffusion. Scientific Reports. 2023;13(1):15666. https://www.nature.com/articles/s41598-023-42891-8.

Pais-Vieira M, Lebedev M, Kunicki C, Wang J, Nicolelis MA. A brain-to-brain interface for real-time sharing of sensorimotor information. Scientific reports. 2013;3(1):1319. https://www.nature.com/articles/srep01319.

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

Paninski L, Pillow J, Lewi J. Statistical models for neural encoding, decoding, and optimal stimulus design. Progress in brain research. 2007;165:493–507.

Park JY, Tsukamoto M, Tanaka M, Kamitani Y. Sound reconstruction from human brain activity via a generative model with brain-like auditory features. arXiv preprint arXiv:230611629. 2023;https://arxiv.org/abs/2306.11629.

Pescatore CR, Zhang H, Hadjinicolaou AE, Paulk AC, Rolston JD, Richardson RM, et al. Decoding semantics from natural speech using human intracranial EEG. bioRxiv. 2025;p. 2025–02. https://www.biorxiv.org/content/10.1101/2025.02.10.637051v1.abstract.

Postolache E, Polouliakh N, Kitano H, Connelly A, Rodolà E, Cosmo L, et al. 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; 2025. p. 1–5. https://ieeexplore.ieee.org/abstract/document/10887735.

Pritchard WS. Psychophysiology of P300. Psychological bulletin. 1981;89(3):506. https://psycnet.apa.org/record/1981-20721-001.

Puffay C, Accou B, Bollens L, Monesi MJ, Vanthornhout J, Francart T, et al. Relating EEG to continuous speech using deep neural networks: a review. Journal of Neural Engineering. 2023;20(4):041003. https://iopscience.iop.org/article/10.1088/1741-2552/ace73f/meta.

Qian X, Wang Y, Fu Y, Sun X, Xue X, Feng J. Joint fmri decoding and encoding with latent embedding alignment. arXiv preprint arXiv:230314730. 2023;https://

arxiv.org/abs/2303.14730.

Qiu W, Huang Z, Hu H, Feng A, Yan Y, Ying R. MindLLM: A Subject-Agnostic and Versatile Model for fMRI-to-Text Decoding. arXiv preprint arXiv:250215786. 2025;https://arxiv.org/abs/2502.15786.

Qiu Y, Liu H, Zhao M. A Review of Brain–Computer Interface-Based Language Decoding: From Signal Interpretation to Intelligent Communication. Applied Sciences. 2025;15(1):392. https://www.mdpi.com/2076-3417/15/1/392.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al.; PmLR. Learning transferable visual models from natural language supervision. 2021;p. 8748–8763. https://proceedings.mlr.press/v139/radford21a.

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI blog. 2019;1(8):9. https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf.

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

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

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

Rossi S, Antal A, Bestmann S, Bikson M, Brewer C, Brockmöller J, et al. Safety and recommendations for TMS use in healthy subjects and patient populations, with updates on training, ethical and regulatory issues: Expert Guidelines. Clinical Neurophysiology. 2021;132(1):269–306. https://www.sciencedirect.com/science/article/pii/S1388245720305149.

Rubin TN, Koyejo O, Gorgolewski KJ, Jones MN, Poldrack RA, Yarkoni T. Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition. PLoS computational biology. 2017;13(10):e1005649. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005649.

S-a R, A-i K, T-s NF. Sanitizing BCI signals: A defense against adversarial attacks. In: Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN) Glasgow, UK; 2020. p. 1–8.

Sato A, Kobayashi I. 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); 2025. p. 53–67. https://aclanthology.org/2025.wraicogs-1.6/.

Schoffelen JM, Oostenveld R, Lam NH, Uddén J, Hultén A, Hagoort P. A 204-subject multimodal neuroimaging dataset to study language processing. Scientific data. 2019;6(1):17. https://www.nature.com/articles/s41597-019-0020-y.

Scotti P, Banerjee A, Goode J, Shabalin S, Nguyen A, Dempster A, et al. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. Advances in Neural Information Processing Systems. 2024;36. https://medarc-ai.github.io/mindeye/.

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

Shehzad A, Yu S, Zhang D, Abid S, Cheng X, Zhou J, et al. BrainGT: Multifunctional Brain Graph Transformer for Brain Disorder Diagnosis. medRxiv. 2024;p. 2024–08.

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

Shen V, Kunanbayev K, Kim DS. Decoding fMRI Data into Captions using Prefix Language Modeling. arXiv preprint arXiv:250102570. 2025;https://arxiv.org/abs/2501.02570.

Shirakawa K, Nagano Y, Tanaka M, Aoki SC, Majima K, Muraki Y, et al. Spurious reconstruction from brain activity. arXiv preprint arXiv:240510078. 2024;https://arxiv.org/abs/2405.10078.

da Silva FL. EEG and MEG: relevance to neuroscience. Neuron. 2013;80(5):1112–1128. https://www.cell.com/neuron/fulltext/S0896-6273(13)00920-3?script=true&mobileUi=0.

Simon AMD, Østergaard J, Bech S, Loquet GSJM. 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; 2022. https://vbn.aau.dk/en/publications/optimal-time-lags-for-linear-cortical-auditory-attention-detectio.

Solon AJ, Lawhern VJ, Touryan J, McDaniel JR, Ries AJ, Gordon SM. Decoding P300 variability using convolutional neural networks. Frontiers in human neuroscience. 2019;13:201. https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2019.00201/full.

Song Y, Liu B, Li X, Shi N, Wang Y, Gao X. Decoding natural images from eeg for object recognition. arXiv preprint arXiv:230813234. 2023;https://arxiv.org/abs/2308.13234.

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

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

Sun J, Li M, Chen Z, Moens MF. Neurocine: Decoding vivid video sequences from human brain activties. arXiv preprint arXiv:240201590. 2024;.

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

Sun J, Li M, Moens MF. Decoding realistic images from brain activity with contrastive self-supervision and latent diffusion. arXiv preprint arXiv:231000318. 2023;.

Sun J, Wang S, Zhang J, Zong C. Towards sentence-level brain decoding with distributed representations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33; 2019. p. 7047–7054.

Sun J, Wang S, Zhang J, Zong C. Neural encoding and decoding with distributed sentence representations. IEEE Transactions on Neural Networks and Learning Systems. 2020;32(2):589–603.

Takagi Y, Nishimoto S. 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; 2023. p. 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.

Takagi Y, Nishimoto S. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. arXiv preprint arXiv:230611536. 2023;https://arxiv.org/abs/2306.11536.

Tang J, LeBel A, Jain S, Huth AG. Semantic reconstruction of continuous language from non-invasive brain recordings. Nature Neuroscience. 2023;26(5):858–866. https://www.nature.com/articles/s41593-023-01304-9.

Tian Z, Quan R, Ma F, Zhan K, Yang Y. BRAINGUARD: Privacy-Preserving Multi-subject Image Reconstructions from Brain Activities. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39; 2025. p. 14414–14422.

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023;https://arxiv.org/abs/2302.13971.

Tripathy K, Markow ZE, Fishell AK, Sherafati A, Burns-Yocum TM, Schroeder ML, et al. Decoding visual information from high-density diffuse optical tomography neuroimaging data. NeuroImage. 2021;226:117516. https://www.sciencedirect.com/science/article/pii/S1053811920310016.

Tripathy K, Markow ZE, Fogarty M, Schroeder ML, Svoboda AM, Eggebrecht AT, et al. Multisensory naturalistic decoding with high-density diffuse optical tomography. Neurophotonics. 2025;12(1):015002–015002. https://pubmed.ncbi.nlm.nih.gov/39850351/.

Tsien JZ. Principles of intelligence: on evolutionary logic of the brain. Frontiers in systems neuroscience. 2016;9:186.

UNESCO. Report of the UNESCO International Bioethics Committee (IBC) on the Ethics of Neurotechnology. Paris, France: UNESCO; 2023.

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

Värbu K, Muhammad N, Muhammad Y. Past, present, and future of EEG-based BCI applications. Sensors. 2022;22(9):3331.

Wang C, Yan H, Huang W, Li J, Wang Y, Fan YS, et al. Reconstructing rapid natural vision with fMRI-conditional video generative adversarial network. Cerebral Cortex. 2022;32(20):4502–4511. https://academic.oup.com/cercor/article/32/20/4502/6515038.

Wang J, Song Z, Ma Z, Qiu X, Zhang M, Zhang Z. Enhancing eeg-to-text decoding through transferable representations from pre-trained contrastive eeg-text masked autoencoder. arXiv preprint arXiv:240217433. 2024;https://arxiv.org/abs/2402.17433.

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

Wang S, Zhang X, Zhang J, Zong C. A synchronized multimodal neuroimaging dataset for studying brain language processing. Scientific Data. 2022;9(1):590.

Wei Z, Dan T, Wu G. Large Connectome Model: An fMRI Foundation Model of Brain Connectomes Empowered by Brain-Environment Interaction in Multitask Learning Landscape. arXiv preprint arXiv:251018910. 2025;.

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

Wiertz S, Boldt J. Evaluating models of consent in changing health research environments. Medicine, Health Care and Philosophy. 2022;25(2):179–189. https://doi.org/10.1007/s11019-022-10074-3, https://doi.org/10.1007/s11019-022-10074-3.

Willett FR, Avansino DT, Hochberg LR, Henderson JM, Shenoy KV. High-performance brain-to-text communication via handwriting. Nature. 2021;593:249–254.

Willett FR, Avansino DT, Hochberg LR, Henderson JM, Shenoy KV. High-performance brain-to-text communication via handwriting. Nature. 2021;593(7858):249–254. https://www.nature.com/articles/s41586-021-03506-2.

Willett FR, Kunz EM, Fan C, Avansino DT, Wilson GH, Choi EY, et al. A high-performance speech neuroprosthesis. Nature. 2023;620(7976):1031–1036. https://www.nature.com/articles/s41586-023-06377-x.

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

Wu H, Li Q, Zhang C, He Z, Ying X. Bridging the Vision-Brain Gap with an Uncertainty-Aware Blur Prior. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025. p. 2246–2257.

Xia W, de Charette R, Oztireli C, Xue JH. Umbrae: Unified multimodal brain decoding. In: European Conference on Computer Vision Springer; 2024. p. 242–259. https://link.springer.com/chapter/10.1007/978-3-031-72667-5_14.

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

Xu L, Xu M, Jung TP, Ming D. Review of brain encoding and decoding mechanisms for EEG-based brain–computer interface. Cognitive neurodynamics. 2021;15(4):569–584.

Xu Y, Liu B, Shao W, Du Y, He S, Zhu T. Cross-Subject Mind Decoding from Inaccurate Representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2025. p. 15066–15075.

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

Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience. 2016;19(3):356–365.

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

Yang YY, Hwang AHC, Wu CT, Huang TR. Person-identifying brainprints are stably embedded in EEG mindprints. Scientific Reports. 2022;12(1):17031. https://doi.org/10.1038/s41598-022-21384-0, https://doi.org/10.1038/s41598-022-21384-0.

Yin C, Ye Z, Li P. Language Reconstruction with Brain Predictive Coding from fMRI Data. arXiv preprint arXiv:240511597. 2024;https://arxiv.org/abs/2405.11597.

Yu S, Chen Z, Wu W, Chen S, Liu Z, Nie J, et al. 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; 2025. p. 6189–6198.

Yuste R, Goering S, et al. Four ethical priorities for neurotechnologies and AI. Nature. 2017;551:159–163.

Zeng B, Li S, Liu X, Gao S, Jiang X, Tang X, et al. Controllable mind visual diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38; 2024. p. 6935–6943. https://ojs.aaai.org/index.php/AAAI/article/view/28519.

Zeng H, Xia N, Qian D, Hattori M, Wang C, Kong W. DM-RE2I: A framework based on diffusion model for the reconstruction from EEG to image. Biomedical Signal Processing and Control. 2023;86:105125. https://www.sciencedirect.com/science/article/abs/pii/S174680942300558X.

Zhang H, Zhao M, Wei C, Mantini D, Li Z, Liu Q. EEGdenoiseNet: a benchmark dataset for deep learning solutions of EEG denoising. Journal of Neural Engineering. 2021;18(5):056057. https://iopscience.iop.org/article/10.1088/1741-2552/ac2bf8/meta.

Zhang X, Xiao M, Meng L. Tiny noise, big mistakes: adversarial perturbations induce errors in brain-computer interface spellers. Natl Sci Rev. 2020;7:1956–1968.

Zhang Y, Ruan H, Yuan Z, Du H, Gao X, Lu J. 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; 2023. p. 1–5. https://ieeexplore.ieee.org/abstract/document/10096819.

Zhang Y, Wang S, Dong X, Yu J, Zong C. Navigating brain language representations: A comparative analysis of neural language models and psychologically plausible models. arXiv preprint arXiv:240419364. 2024;.

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

Zheng R, Sun L. Llm4brain: Training a large language model for brain video understanding. arXiv preprint arXiv:240917987. 2024;https://arxiv.org/abs/2409.17987.

Zhou J, Xie Y, Fan C, Wang H, Lv Z, Tao L. DHGCN: Dual HyperGraph Convolutional Network for EEG-Based Auditory Attention Detection. In: Proceedings of the 33rd ACM International Conference on Multimedia; 2025. p. 612–620. https://doi.org/10.1145/3746027.3755579.

Zhou M, Tian C, Cao R, Wang B, Niu Y, Hu T, et al. Epileptic seizure detection based on EEG signals and CNN. Frontiers in neuroinformatics. 2018;12:95. https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2018.00095/full.

Zhou M, Gong Z, Dai Y, Wen Y, Liu Y, Zhen Z. A large-scale fMRI dataset for human action recognition. Scientific Data. 2023;10(1):415. https://www.nature.com/articles/s41597-023-02325-6.

Zhu D, Bieger J, Garcia Molina G, Aarts RM. A survey of stimulation methods used in SSVEP-based BCIs. Computational intelligence and neuroscience. 2010;2010(1):702357. https://onlinelibrary.wiley.com/doi/full/10.1155/2010/702357.

Zhu S, Ye Z, Ai Q, Liu Y. EEG-ImageNet: An Electroencephalogram Dataset and Benchmarks with Image Visual Stimuli of Multi-Granularity Labels. arXiv preprint arXiv:240607151. 2024;https://arxiv.org/abs/2406.07151.

Zhu X, Chai Y, Li R, Lan M, Gao L. Decoding the Moving Mind: Multi-Subject fMRI-to-Video Retrieval with MLLM Semantic Grounding. bioRxiv. 2025;p. 2025–04. https://www.biorxiv.org/content/10.1101/2025.04.07.647335v1.abstract.

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

de Zuazo X, Verbeni V, Ku LC, Arrieta E, Barrena A, Klimovich-Gray A, et al. # neural2speech: Decoding Speech and Language from the Human Brain. In: Proc. IberSPEECH 2024; 2024. p. 256–260. https://www.isca-archive.org/iberspeech_2024/dezuazo24b_iberspeech.html.

Zuboff S. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. London, UK: Profile Books; 2019.