Proceedings of ICEIPI 2025 Symposium: Understanding Religious Identity in Educational Contexts
DOI: 10.54254/2753-7048/2025.ND25514

A Review of Multimodal Brain Language Decoding

Yixuan Tang

TCE International School, Kamphaeng Phet, Thailand
savannah.tang10@gmail.com

Abstract. Brain language decoding is a frontier field bridging cognitive neuroscience and artificial intelligence, which aims to reconstruct language cognition from neural signals. This review summarizes recent progress in multi-modal decoding approaches. Methodologically, it highlights advancements in integrating neural acquisition techniques such as fMRI, EEG, and ECoG with deep neural network modeling. Theoretically, it explores core issues including signal spatiotemporal characteristics, feature alignment, and cross-modal generalization. At the application level, the paper assess the translational potential of this technology in clinical language disorder interventions and intelligent brain-computer interfaces. To address challenges such as individual variability and limited high-quality data, it discusses strategies including universal decoding frameworks and adaptive learning algorithms. This study also reflects on ethical concerns related to neurotechnology. Together, this work outlines a systematic framework and technical roadmap for advancing brain-language decoding in both fundamental research and real-world applications.

Keywords: multimodal, brain language decoding, cognitive neuroscience, artificial intelligence

1. Introduction

Language is one of the most complex and representative higher cognitive functions in humans. Understanding how the human brain encodes, processes and generates language is not only crucial for revealing the neural mechanisms of language cognition, but also has important application value in artificial intelligence, human-computer interaction, clinical rehabilitation and other fields. With the rapid development of neuroimaging technology and machine learning methods, Brain Language Decoding (BLD), as a cutting-edge field of research at the intersection of cognitive neuroscience and artificial intelligence, is gradually realizing the goal of restoring linguistic content from neural signals in the brain.

In recent years, Brain Language Decoding has made important breakthroughs in revealing the spatial and temporal dynamics of language processing by integrating multimodal neural signals (including fMRI, EEG, and ECoG) and combining with computational modeling methods. The field has gradually established a systematic research framework of "neural signal acquisition-higher-order feature characterization-semantic space mapping", and its paradigm development has shown two significant trends: in terms of experimental design, research has shifted from the early discrete lexical stimulus paradigm [1,2] to continuous multimodal linguistic narrative scenes, and to

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2025.ND25514

continuous multimodal narrative scenes [3,4], significantly improving the ecological validity of the research; at the methodological level, the research paradigm has undergone a shift from linear interpretable models to deep learning frameworks, with architectures such as Convolutional Neural Networks, Long and Short-Term Memory Networks, and Transformer effectively improving the extraction of complex semantic features through distributed representation learning [5,6]. However, the field still faces several key challenges: first, individual neural representation differences and multimodal data heterogeneity lead to a significant degradation of model cross-subject generalization performance (by 35%-50%); second, the black-box nature of deep learning models constrains the theoretical elucidation of neurocognitive mechanisms; furthermore, as semantic decoding accuracy surpasses 70% [7], the neural privacy leakage risk [8] has transformed from a theoretical discussion to a real-world problem, and there is an urgent need to build an interdisciplinary ethical governance system.

This paper systematically reviews the research progress in the field of brain-language decoding, focusing on the four major modalities of text, speech, image, and video, and provides an in-depth analysis of the technical routes, key breakthroughs, and existing challenges in multimodal decoding. This review not only provides systematic knowledge for understanding the neural mechanisms of multimodal language processing, but also lays a methodological foundation for developing the next generation of brain-computer interface technologies, and points out the key research directions in the future, such as individual difference correction and model interpretability enhancement.

2. Current status and progress of multimodal language brain decoding research

Natural language understanding relies on the dynamic integration of multisensory inputs, necessitating decoding paradigms that transcend single-modality constraints. This study employs a multimodal fusion framework, combining the spatial precision of fMRI, the temporal resolution of EEG, and the representational capacity of deep neural networks, to investigate the brain's encoding of text, speech, image, and video stimuli. We aim to uncover modality-specific representation patterns and integration mechanisms within language-related neural pathways, and to evaluate the limits of current decoding models in processing complex, naturalistic language inputs.

2.1. Text modality: decoding linguistic semantics from reading tasks

As the most maturely developed modality in linguistic brain decoding research, the core goal of the textual modality is to reveal the mapping relationship between the neural responses generated by individuals during the reading process and linguistic semantics. Early experimental paradigms mainly used highly controllable task designs, such as presenting static linguistic stimuli such as words or phrases, and combined with neuroimaging techniques such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), or electrocortical electroencephalography (ECoG). Mitchell et al. pioneered the field by proposing a semantic feature space based on lexical co-occurrence statistics, and successfully predicted the fMRI activation patterns of neologisms in cortical regions. This study laid the foundation for neurosemantic modeling, and a series of subsequent studies further extended the coverage of the semantic space and the expressive power of contextual information, employing methods such as Ridge regression with GloVe word vectors to improve Top-5 decoding accuracy to 71%, much higher than the randomized baseline level of 16.7% [9,10]. As the technology has evolved, researchers have gradually moved away from controlled tasks to more naturalistic language comprehension scenarios for more ecologically valid neural response data.

DOI: 10.54254/2753-7048/2025.ND25514

A key breakthrough in text modality research comes from Huth et al., who constructed the first whole-brain-scale "semantic topography" through language listening experiments in natural contexts [3]. Based on fMRI responses to more than 2,000 high-frequency words, the study systematically revealed the distributed encoding mechanism of semantic information in the cerebral cortex.

With the development of deep learning, pre-trained language models (e.g., BERT and GPT) have been introduced into brain-language research for a more nuanced portrayal of the correspondence between linguistic representations and brain activity, as shown in Figure 1. Toneva and Wehbe and Caucheteux and King found that the mid-level representations of the Transformer model (layers 4 to 6) were highly correlated with early language processing areas (e.g., the superior temporal sulcus), whereas their high-level representations (layers 9 to 11) showed stronger correlations (correlation coefficients up to r = 0.67) with higher cognitive areas such as the prefrontal cortex and the default mode network [6,11]. The Brain-Score framework developed by Schrimpf et al. further validated the ability of deep language models to predict activation patterns in language-related brain regions, significantly outperforming traditional methods, advancing our understanding of hierarchical neural mechanisms of language processing [5].

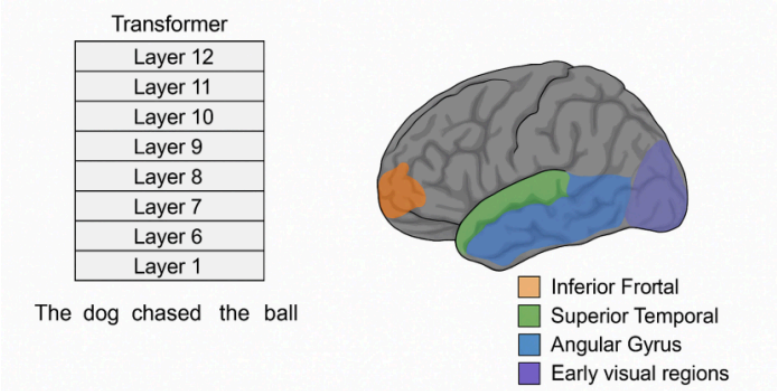


Figure 1. Alignment of language model hierarchical representations with cortical language processing areas

Several high-quality datasets have been constructed to support this research: the Narratives dataset (built by the Huth lab) contains fMRI recordings from a natural story listening task; the Pereira 2018 dataset contains 180 semantically rich sentences designed for cross-subject semantic alignment; and the BrainBench dataset integrates multiple modalities of EEG, fMRI, and ECoG to support comparative analyses of English and Chinese, fMRI, and ECoG in multiple modalities, and supports bilingual comparative analysis. The key features of these three types of datasets are summarized in Table 1.

DOI: 10.54254/2753-7048/2025.ND25514

Table 1. Overview of representative brain language decoding datasets

Dataset Name	Language	Modality	Number of Subjects	Features
Narratives (Huth)	English	fMRI	7	Naturalistic story stimuli; used to construct whole-brain semantic maps
Pereira 2018	English	fMRI + Text Embedding	16	Sentence-level semantic classification; supports cross-subject decoding analysis
BrainBench (2021)	English / Chinese	EEG / fMRI / ECoG	Various	Multimodal, bilingual benchmark; enables cross-task comparisons

Overall, text modality research has developed from early lexical-level semantic decoding to chapter-level semantic reconstruction. This direction has formed a more systematic research framework by constructing semantic maps, sharing neural embedding space, and introducing deep neural network representation analysis and multimodal data fusion techniques. This not only expands our theoretical understanding of the neural mechanisms of language, but also provides solid technical support for clinical applications such as brain-computer interface design and language disorder rehabilitation. However, the modeling of chapter structure, the accuracy of context alignment, and the ecological adaptability of the decoding system are still the key bottlenecks for further breakthroughs in the future.

2.2. Speech modality: decoding linguistic information from natural auditory inputs

As a key input channel for language understanding, brain decoding research in speech modality focuses on extracting language-related information from auditory-evoked neural activities. This research direction faces two core challenges: first, speech signals are characterized by millisecond-level dynamics, which requires neural decoding with extremely high temporal resolution; second, speech processing involves multilevel neural pathways from the primary auditory cortex to high-level linguistic regions, which requires models that can efficiently model the hierarchical structure and temporal dependence. These features make speech modality an important window to understand the neural transformation mechanism of "perception-cognition-language", but at the same time, it also puts forward very high requirements on signal acquisition and modeling methods.

In early studies on phoneme levels, Chang's team and Mesgarani et al. systematically mapped the neural representation of phonemes in the primary auditory cortex of the superior temporal gyrus by using electrocortical electroencephalography (ECoG) [12,13]. They found that different phoneme classes (e.g., vowels vs. consonants) evoked significantly distinguishable cortical neural responses, and that these distinctions could be detected within 50-150 ms after speech stimulation, with classification accuracies of more than 80% for ECoG signals. These findings establish the neural basis for phoneme-level neural decoding and provide theoretical support for subsequent fast speech recognition and synthesis systems.

Speech neural decoding has realized the evolution from the acoustic spectrum to natural language output in the reconstruction path. Pasley et al. first proposed a model to reverse map the speech spectrum from fMRI and ECoG signals, and verified the feasibility of reconstructing speech based on neural activity (with 60% accuracy in semantic recognition) [2]. Subsequently, Akbari et al. introduced a CNN-LSTM architecture with vocoder technology, which significantly improved the perceptual quality and phoneme recognition rate (>80%) of speech reconstruction [14]. Angrick et al. integrated attention mechanisms into an end-to-end speech synthesis model, achieving more than

DOI: 10.54254/2753-7048/2025.ND25514

15% improvement in PESQ metrics [15]. Sun et al. further realized the complete process from brain signal to readable text, constructed an end-to-end decoding system of "brain signal → acoustic features → linguistic text", and achieved a phoneme recognition rate of 73.1% and a word recognition rate of 42.5% under the condition of only 30 minutes of training samples [16].

In recent years, researchers have gradually explored the correspondence between pre-trained speech models and neural activities in the brain, and Alain et al. systematically analyzed the relationship between different levels of representations of the wav2vec 2.0 model and the neural responses of brain regions, and found that the middle level representations were highly synchronized with primary auditory regions such as the Heschl's gyrus (r=0.58), whereas the higher level representations better corresponded to the superior temporal gyrus, middle temporal gyrus, and inferior frontal cortex. gyrus and inferior frontal cortex (r=0.63) [17]. While Berezutskaya et al. showed that wav2vec representations improved the interpretation of neural signals by 42% compared to traditional MFCC features, providing strong support for the development of speech models that are more consistent with neural mechanisms [18].

Cross-language and cross-speaker generalization are the key directions to promote the practicality of speech brain decoding. Yang et al. developed a multilingual decoding framework that achieves cross-linguistic classification of syllables and semantics in Chinese and English bilingual subjects (with an accuracy of 82%), and their cross-speaker pre-training strategy significantly improves the phoneme recognition performance of new subjects (with an accuracy of about 80%) [19]. Angrick et al. improved the phoneme recognition performance of Chinese and English subjects through small-sample fine-tuning, and their cross-speaker pre-training strategy improved the phoneme recognition performance of Chinese and English bilinguals.), on the other hand, improved the performance of the decoding model on new individuals through small-sample fine-tuning (up to 18-25% improvement), providing a technical path for the construction of a generalized neural speech decoding system[20].

Speech modality research has initially established a systematic framework from acoustic feature recognition to semantic reconstruction and achieved a series of breakthrough results. However, several key challenges remain: first, there is a natural contradiction between the high temporal properties of speech signals and the spatiotemporal resolution of neuroimaging means, which puts pressure on the temporal modeling of the model; second, phonemes with similar acoustic features have significant overlap in cortical representations, which leads to blurring of the classification boundaries (with an average confounding error as high as 35%); furthermore, speech processing involves the synergy of multimodal brain regions, including the motor cortex (articulatory preparation), prefrontal lobe (attentional control), and limbic system (emotional processing), which introduce significant noise interference in actual decoding.

Overall, speech modality provides an important window into understanding how the human brain processes continuous speech input. From the early days of phoneme-level classification and spectral reconstruction to today's cross-linguistic modeling and end-to-end natural language generation, the field has shown great potential in terms of underlying cognitive mechanisms, modeling techniques, and translation to clinical applications. With the continuous evolution of high-resolution neural signal acquisition techniques, deep neural networks, and multimodal modeling strategies, speech brain decoding is expected to play a greater role in practical scenarios such as neurospeech interfaces, aphasia rehabilitation, and silent speech communication.

DOI: 10.54254/2753-7048/2025.ND25514

2.3. Image modality: language semantic representation and decoding in cross-sensory input

Image modality plays a unique role in language brain decoding research. Typical experimental tasks include image naming and visual semantic judgment, which are widely used to investigate the coupling between the visual cortex and language networks. Studies have shown that when participants process text-image information or describe images, visual pathways (such as the ventral visual pathway) establish specific functional connections with language-related regions (such as the angular gyrus, middle temporal gyrus, and prefrontal cortex), providing a neural basis for semantic generation and representation.

Research on neural representation mechanisms has revealed the distribution characteristics of visual features in the cerebral cortex and their high consistency with artificial models. Wen et al. found that compared with traditional low-level features (such as Gabor), VGG and other deep models exhibit stronger correlations with neural activity in the prefrontal and parietal cortices ($r = 0.68$ vs. $r = 0.32$) [21]; Khaligh-Razavi and Kriegeskorte further noted that the semantic space of deep networks shows significant structural alignment ($r > 0.7$) with the temporal visual cortex (e.g., IT and LOC) ($r > 0.7$) [22]. Yamins et al. constructed a visual computation model with a response correlation coefficient of 0.75 to the inferior temporal cortex (IT), establishing the core paradigm for "artificial model-neural representation" alignment research. The embedding spaces of models like CLIP have also been shown to effectively map to language networks, serving as a novel intermediary representation for multimodal brain decoding [23].

Image-induced language generation is one of the key research directions in image modality decoding. Horikawa and Kamitani were the first to map fMRI activations to GloVe/BERT embedding spaces to predict image-related semantic labels (accuracy > 65%) [24]; Carlson et al. built a complete decoding pipeline using CLIP representations as an intermediary: "visual stimulus → neural activity → semantic vector → language generation," achieving a semantic matching rate of 52.3% in image description tasks [4]. These studies validate the cross-modal neural pathway from images to language output, providing theoretical and methodological foundations for realizing "image → brain → language" brain-computer interfaces.

In the field of multimodal neural representation research, Caucheteux et al. recently demonstrated that when visual images are semantically consistent with accompanying text, the middle temporal gyrus and posterior cingulate cortex exhibit significant synergistic activation ($p<0.001$) and enhanced functional connectivity ($r=0.72$), confirming the critical role of cross-modal semantic consistency in neural integration [25]. This finding is consistent with the results of a bilingual study by He et al., which found that fMRI patterns induced by the same image stimuli in different language contexts exhibit high consistency in semantic processing regions of the temporal and parietal lobes (intra-group correlation coefficient ICC = 0.82), revealing that image semantic representations possess cross-linguistic stability [26]. These findings provide important neuroscientific evidence for the development of brain decoding systems that are universal across languages. Current brain decoding research on image modalities primarily relies on three standardized datasets: the Natural Scenes Dataset (NSD) provides a benchmark for analyzing neural response consistency through high-replication fMRI recordings; the ImageNet-fMRI dataset supports research on the neural mechanisms of object recognition; and the BOLD5000 dataset, with its rich image categories (over 5,000) and diverse semantic content, is an ideal choice for constructing complex semantic neural representation models. The establishment of these datasets has significantly enhanced the comparability and reproducibility of cross-laboratory studies.

Despite the obvious cross-modal decoding advantages of image modalities, three major challenges remain: first, there are significant subjective differences in individuals' semantic

DOI: 10.54254/2753-7048/2025.ND25514

descriptions of images, resulting in an average semantic overlap rate of only approximately 45–60%, which affects semantic labeling and model training; second, image-related information is highly distributed in the brain, involving primary visual areas (V1–V4), mid-temporal semantic areas, and attention control networks, increasing the modeling complexity of decoding models; third, dynamic visual stimuli (such as videos) require millisecond-level temporal resolution to sample neural responses, while current fMRI and other techniques cannot meet the temporal resolution requirements for modeling. These technical and theoretical bottlenecks urgently require breakthroughs through the integration of cross-modal modeling strategies and neural enhancement technologies.

In summary, the image modality provides an important pathway for understanding the perception-language conversion mechanism. Its research, through the synergistic application of multimodal embedding spaces, deep generative models, and cross-modal alignment strategies, continues to drive brain-to-language decoding toward more complex, natural, and multicontextual directions.

2.4. Video modality: multimodal semantic integration and brain-to-language decoding in temporal dynamics

In natural communication environments, language rarely appears in isolation but is often synchronously intertwined with visual, auditory, and emotional modalities to form a complete context. Video, as a complex dynamic input integrating visual images, speech intonation, linguistic structure, and narrative information, serves as an ideal paradigm for simulating natural language scenarios. Research indicates that compared to static modalities such as text or images, video stimuli activate broader and more sustained brain regions, particularly higher-order cortical areas such as the parietal lobe, prefrontal cortex, and default mode network (DMN) [27]. Therefore, the video modality not only offers experimental advantages with high ecological validity but also provides a crucial window into core cognitive processes such as discourse-level language construction, multisensory integration, and event understanding. Current video modality research primarily employs three paradigms, including free viewing tasks, alignment modeling tasks, and semantic decoding/generation tasks.

In terms of technical approaches, researchers typically employ multi-stream encoder structures to extract visual (e.g., ResNet or I3D), auditory (e.g., wav2vec), and linguistic (e.g., BERT/GPT) features from videos, followed by Transformer or temporal attention mechanisms to construct a continuous semantic space, and then use classifiers or language generators to predict events, characters, or sentences. Vodrahalli et al. conducted an fMRI study based on videos and subtitles, and were the first to construct a cross-modal mapping model of "visual/language embeddings-neural activity," achieving high-accuracy predictions of verb/noun categories (55.6%) [28]; Zellers et al. integrated GPT-3 representations, proposed a decoding path of "semantic graph → neural response → natural language generation," achieving significant advantages in story event prediction tasks [29]. Caucheteux and King further analyzed neural patterns across multiple movie stimuli, finding that GPT representations highly aligned with brain activity in the left prefrontal cortex ($r = 0.64$), providing quantitative evidence for cross-modal neural alignment [11].

The construction of standardized video neural datasets provides critical resources for cross-laboratory collaborative research. The Narratives Dataset collects whole-brain fMRI data from over 300 participants watching natural films, accompanied by frame-level semantic annotations, making it the first benchmark platform supporting triple alignment of neural signals, video content, and language descriptions. The Sherlock Movie Dataset innovatively introduces film narrative structure

DOI: 10.54254/2753-7048/2025.ND25514

analysis, providing event segmentation, character labels, and timeline alignment, making it particularly suitable for studying neural-language mapping mechanisms in complex narratives. These datasets, through shared protocols and standardized task designs, significantly enhance the comparability and reproducibility of research results, accelerating the field's transition from exploratory modeling to reproducible science.

Although video modalities demonstrate rich cognitive value and modeling potential, they still face numerous challenges. First, video content exhibits high temporal dynamics, while brain imaging techniques such as fMRI have sampling rates (e.g., TR=2s) that struggle to capture frame-level neural changes, limiting temporal precision. Second, individual differences among participants in semantic construction and scene understanding complicate the development of generalizable models. Third, multimodal information in videos, such as images, speech, subtitles, and other multimodal information in videos, may exhibit temporal and semantic asynchrony, leading to "semantic drift" issues and exacerbating the difficulty of modal alignment in decoding tasks. These technical and theoretical bottlenecks are the core challenges that current video modality brain-language decoding research urgently needs to address.

3. Influencing factors and research challenges in brain language decoding

In natural language communication, linguistic information is typically integrated with multimodal information such as visual, auditory, and emotional cues to form a complete semantic understanding context. As a dynamic composite stimulus that integrates visual scenes, speech signals, linguistic structures, and narrative cues, video can effectively simulate real-world language communication environments. Neuroscience research shows that compared with static text or image stimuli, video input can induce broader and more sustained brain activation patterns, particularly significantly enhancing neural activity in the parietal association cortex, prefrontal cortex, and default mode network (DMN) [27]. This characteristic makes the video modality an important experimental paradigm for studying the neural mechanisms of multimodal information integration and contextual understanding.Current brain neuroscience research primarily employs three experimental paradigms, including free viewing paradigm, subtitle synchronization paradigm, and semantic reconstruction paradigm.

At the methodological level, current research mainly uses multi-stream encoder architectures to extract visual (e.g., ResNet/I3D) and speech (e.g., wav2vec), and language (e.g., BERT/GPT) features from videos, and utilizes temporal attention mechanisms or Transformer structures to achieve cross-modal semantic integration. Vodrahalli et al. pioneered the matching of multimodal semantic vectors with fMRI signals, successfully predicting verb and noun categories in videos, thereby validating the effectiveness of multimodal features [28]. Subsequent studies have further refined this technical approach: Zellers et al. introduced GPT-3 to build a generative model from video semantics to natural language, demonstrating superior performance to traditional methods in narrative reconstruction tasks [29]; Caucheteux and King conducted a quantitative analysis and found that GPT representations are significantly correlated with neural activity in the left prefrontal cortex ($r = 0.64$), providing direct evidence for the correspondence between language model representations and semantic processing in the brain [11].

The establishment of standardized datasets provides an important foundation for neural decoding research in the video modality. The Narratives Dataset integrates fMRI data from over 300 participants watching films, supplemented with frame-level semantic annotations, achieving the first three-dimensional spatiotemporal alignment of neural activity, video content, and language descriptions. The Sherlock Movie Dataset further provides fine-grained event segmentation,

DOI: 10.54254/2753-7048/2025.ND25514

character recognition, and temporal annotation information, creating conditions for research on the neural representation of narrative structure. The innovations of these two datasets in experimental paradigm design, data quality control, and open sharing not only improve the comparability of research results but also lay the foundation for the repeatability of methodologies.

Although significant progress has been made in brain-to-video decoding research, several key challenges remain. The performance of brain-language decoding models is primarily constrained by three key factors: first, the choice of neural signal acquisition technology directly impacts the upper limit of decoding performance. The fMRI offers millimeter-level spatial resolution (1–3 mm), but its temporal resolution is typically limited to 2–3 seconds, making it difficult to capture the rapid dynamic processes of language processing; electroencephalography (EEG) has millisecond-level temporal resolution, but is susceptible to artifacts such as muscle activity and eye movements; while cortical EEG (ECoG) can combine temporal and spatial resolution (spatial resolution of approximately 1 cm and temporal resolution of approximately 5 ms), its invasive nature makes it difficult to apply widely in healthy populations [30].

Second, the hierarchical nature of language processing significantly increases decoding difficulty. Neuroimaging studies indicate that primary speech perception primarily relies on auditory cortex regions such as the superior temporal gyrus, while higher-level semantic integration and syntactic processing require large-scale network coordination across multiple brain regions such as the prefrontal cortex [31]. This hierarchical neural representation mechanism requires decoding models to simultaneously process neural activity across different time scales and spatial ranges, thereby greatly increasing the complexity of data collection and computational modeling.

Finally, existing methodologies still have obvious limitations. Although deep learning models have achieved significant improvements in decoding accuracy, their inherent “black box” nature severely hinders theoretical explanations of neural representation mechanisms [5]. Additionally, models exhibit limited generalization performance across languages and cultural contexts, and establishing a universally applicable semantic neural representation framework remains a core scientific challenge in this field.

4. Applications and prospects of brain-language decoding research

4.1. Applications

Brain-language decoding technology has demonstrated significant application value in multimodal modeling and real-time prediction. In the clinical medicine field, this technology provides novel rehabilitation methods for patients with language disorders such as aphasia [32]; in education, neurofeedback-based language comprehension assessment systems hold promise for personalized teaching adaptation [33]; in human-computer interaction, its integration with VR/AR technologies can enable the development of more context-aware intelligent systems [34]; and in basic research, this technology provides a new research paradigm for revealing the neural mechanisms of language processing [3]. It is worth noting that with the development of technology, it is necessary to establish strict ethical review systems and embed compliance mechanisms in the process of technology development and application to ensure that its development complies with ethical norms and prevents potential risks.

DOI: 10.54254/2753-7048/2025.ND25514

4.2. Future prospects

Future brain-to-language decoding research will focus on the following key directions. First, in terms of decoding scope, it is necessary to develop long-sequence modeling methods based on architectures such as Transformer-XL to achieve semantic parsing from the lexical to the paragraph level, thereby enhancing language understanding capabilities in natural contexts [35]. Second, in terms of model generalization, a general decoding framework should be constructed through meta-learning and cross-modal transfer learning to address the issue of neural representation differences in multi-language and multi-cultural contexts [36]. Third, in terms of technology integration, we have integrated cross-modal pre-training models such as CLIP to establish a unified representation system for vision, hearing, and language, enhancing the ecological validity of decoding systems [29]. Additionally, it is crucial to improve model interpretability by introducing neural oscillation features or developing neural-symbol hybrid models to enhance the compatibility of decoding results with cognitive theories [5]. At the application level, it is necessary to optimize lightweight algorithms and portable acquisition devices (such as fNIRS) to promote the practical application of real-time brain-language interfaces in clinical rehabilitation and human-computer interaction. It is worth noting that with the improvement of decoding accuracy, it is necessary to simultaneously establish and improve neuroethical norms to balance technological innovation and privacy protection [8].

5. Conclusion

This paper systematically reviews the research progress in the field of multimodal brain-language decoding, focusing on decoding methods, neural representation mechanisms, and core scientific issues in four modalities: text, speech, image, and video. The study reveals three major challenges: the spatio-temporal characteristics of neural signals, the hierarchical complexity of language processing, and the lack of model interpretability and cross-subject generalization ability. Based on this, this paper proposes four key directions for future research: the construction of a general decoding architecture, multimodal deep integration, the development of real-time application technologies, and the improvement of ethical frameworks.

Although this study provides a comprehensive review of mainstream technical approaches, there are still shortcomings in the exploration of the application potential of emerging methods such as diffusion models and cross-cultural differences in neural representations. Follow-up research can deepen these directions through multi-center collaborative experiments and the construction of cross-modal datasets. As a bridge connecting cognitive neuroscience and artificial intelligence, breakthroughs in multimodal brain-language decoding technology will profoundly impact clinical rehabilitation, human-machine interaction, and brain function assessment, providing new research paradigms for understanding the essence of human language cognition.

References

[1] Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. Science, 320(5880), 1191–1195. https: //doi.org/10.1126/science.1152876

[2] Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., & Chang, E. F. (2012). Reconstructing speech from human auditory cortex. PLoS Biology, 10(1), e1001251. https: //doi.org/10.1371/journal.pbio.1001251

[3] Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532, 453–458. https: //doi.org/10.1038/nature17637

DOI: 10.54254/2753-7048/2025.ND25514

[4] Carlson, T. A., Goddard, E., & Woolgar, A. (2023). Decoding complex scene representations from the human brain using multimodal deep learning. Nature Human Behaviour, 7, 212–222. https: //doi.org/10.1038/s41562-022-01488-9
[5] Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. PNAS, 118(45), e2105646118. https: //doi.org/10.1073/pnas.2105646118
[6] Toneva, M., & Wehbe, L. (2022). Combining machine and brain representations of language. Nature Machine Intelligence, 4, 112–122. https: //doi.org/10.1038/s42256-021-00414-z
[7] Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Geva, R., Kim, J. W., Popham, S., Dunlap, K., Brennan, J. R., Huth, A. G., & Hasson, U. (2022). Thinking ahead: Prediction in context as a keystone of language in humans and machines. bioRxiv. https: //doi.org/10.1101/2022.01.29.478314
[8] Martin, C., Wehbe, L., & Rao, R. P. N. (2019). Neural privacy: How machine learning could enable privacy violations in the brain. Nature Neuroscience, 22, 1645–1646. https: //doi.org/10.1038/s41593-019-0492-1
[9] Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., & Mitchell, T. M. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. NeuroImage, 62(1), 451–463. https: //doi.org/10.1016/j.neuroimage.2012.04.048
[10] Anderson, A. J., Zinszer, B. D., & Raizada, R. D. S. (2016). Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. NeuroImage, 128, 44–53. https: //doi.org/10.1016/j.neuroimage.2015.12.035
[11] Caucheteux, C., & King, J.-R. (2022). GPT-2's activations predict the degree of semantic comprehension in the human brain. Cognitive Neuroscience, 13(3–4), 147–154. https: //doi.org/10.1080/17588928.2022.2103653
[12] Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. Nature Neuroscience, 13(11), 1428–1432. https: //doi.org/10.1038/nn.2641
[13] Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science, 343(6174), 1006–1010. https: //doi.org/10.1126/science.1245994
[14] Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. Scientific Reports, 9, 874. https: //doi.org/10.1038/s41598-018-37359-z
[15] Angrick, M., Samek, W., & Meinecke, F. C. (2019). Speech synthesis from ECoG using densely connected 3D convolutional neural networks with vocoder representations. Journal of Neural Engineering, 16(3), 036019. https: //doi.org/10.1088/1741-2552/ab0c46
[16] Sun, P. F., Moses, D. A., Metzger, S. L., Anumanchipalli, G. K., & Chang, E. F. (2023). Decoding speech from non-invasive brain signals with self-supervised learning. Nature Neuroscience, 26, 1232–1242. https: //doi.org/10.1038/s41593-023-01358-4
[17] Alain, C., Shahin, A., He, Y., & Alain, C. (2022). Neural tracking of hierarchical speech representations in the brain with wav2vec 2.0. NeuroImage, 261, 119478. https: //doi.org/10.1016/j.neuroimage.2022.119478
[18] Berezutskaya, J., Alex, J., Cohen, M., & Chang, E. F. (2020). Wav2vec representation improves interpretability of brain responses to speech. bioRxiv. https: //doi.org/10.1101/2020.10.26.354050
[19] Yang, Y., Sun, P., Zheng, Y., Chen, M., & Wang, Y. (2022). Cross-lingual and cross-speaker speech decoding from ECoG using transfer learning. IEEE Transactions on Biomedical Engineering, 69(8), 2430–2440. https: //doi.org/10.1109/TBME.2022.3164074
[20] Angrick, M., et al. (2021). Subject-independent decoding of speech from intracranial brain activity using transfer learning. Frontiers in Neuroscience, 15, 682639. https: //doi.org/10.3389/fnins.2021.682639
[21] Wen, H., Shi, J., Chen, W., & Liu, Z. (2017). Deep residual network predicts cortical representation and organization of visual features for rapid categorization. Scientific Reports, 7, 16214. https: //doi.org/10.1038/s41598-017-16262-3
[22] Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Computational Biology, 10(11), e1003915. https: //doi.org/10.1371/journal.pcbi.1003915
[23] Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. PNAS, 111(23), 8619–8624. https: //doi.org/10.1073/pnas.1403112111
[24] Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. Nature Communications, 8, 15037. https: //doi.org/10.1038/ncomms15037

DOI: 10.54254/2753-7048/2025.ND25514

[25] Caucheteux, C., Posada, A., & King, J. R. (2023). Visual and linguistic modalities converge on a shared neural code for meaning. Nature Communications, 14, 2393. https: //doi.org/10.1038/s41467-023-37977-7
[26] He, Y., Chen, M., & Chen, Y. (2020). Cross-linguistic consistency in neural representation of image semantics. Journal of Cognitive Neuroscience, 32(12), 2447–2460. https: //doi.org/10.1162/jocn_a_01506
[27] Chen, P.-H. C., et al. (2017). A reduced-dimension fMRI shared response model. Advances in Neural Information Processing Systems, 30, 460–468.
[28] Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Chen, J., Yong, E., Honey, C., & Arora, S. (2017). Mapping between fMRI responses to movies and their natural language annotations. NeuroImage, 180(Pt A), 223–231. https: //doi.org/10.1016/j.neuroimage.2017.06.042
[29] Zellers, R., Holtzman, A., Rashkin, H., & Choi, Y. (2021). PIGLeT: Language grounding through neuro-symbolic interaction in a 3D world. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 1255–1267. https: //aclanthology.org/2021.acl-long.100/
[30] Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: Decoding spoken sentences from phone representations in the brain. Frontiers in Neuroscience, 9, 217. https: //doi.org/10.3389/fnins.2015.00217
[31] Fedorenko, E., Behr, M. K., & Kanwisher, N. (2016). Functional specificity for high-level linguistic processing in the human brain. PNAS, 113(41), E6050–E6059. https: //doi.org/10.1073/pnas.1603312113
[32] Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., & Chang, E. F. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. New England Journal of Medicine, 385(3), 217–227. https: //doi.org/10.1056/NEJMoa2027540
[33] Yoo, A. B., Lee, C., & Lim, H. (2021). EEG-based adaptive learning system for second language learners: A neurofeedback approach. Computers & Education, 161, 104072. https: //doi.org/10.1016/j.compedu.2020.104072
[34] Chen, X., Wang, X., Wang, Y., & Liu, J. (2022). Multimodal neural interfaces for immersive VR: Brain decoding and neurofeedback in human-computer interaction. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30, 105–116. https: //doi.org/10.1109/TNSRE.2021.3135547
[35] Jain, S., & Huth, A. G. (2020). Incorporating context into language encoding models for fMRI. Advances in Neural Information Processing Systems, 33, 5796–5806.
[36] Xu, Y., Zhang, Y., Guo, X., & Li, P. (2021). Cross-lingual brain decoding with transfer learning. NeuroImage, 225, 117463. https: //doi.org/10.1016/j.neuroimage.2020.117463