Cross-Subject Universal Neural Decoding Methods for Multi-tasking and Subject Data Migration

Diwei Wu

Department of Integrated Program for Neuroscience

McGill University

August 17, 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Neuroscience (Thesis)

Table of Contents

1. Introduction ····· 6

1.1 Statement of Contribution ····· 6

1.2 Neural Decoding ····· 6

1.3 Visual Neural Decoding Task ····· 8

1.3.1 Category Decoding ····· 9

1.3.2 Reconstruct Decoding ····· 9

1.3.3 Caption Decoding ····· 10

1.4 Cross-subject Data Alignment ····· 11

1.4.1 Anatomical Alignment ····· 12

1.4.2 Functional Alignment ····· 13

1.5 GPT-based Large Language Model ····· 15

1.6 Development in Universal Neural Decoding ····· 16

1.7 Research Content and Significance ····· 18

2. Multi-task cross-subject universal decoding model based on GPT ····· 20

2.1 Introduction ····· 20

2.2 Experiment Data ····· 22

2.3 Model ····· 23

2.3.1 Prompt Design ····· 24

2.3.2 Word Embedding Module ····· 25

2.3.3 Position Embedding Module ····· 26

2.3.4 Vision Encoder Module ····· 26

2.3.5 Network Encoder Module ····· 28

2.3.6 GPT Decoder Module ····· 29

2.4 Decoding experiment results ····· 30

2.4.1 Category Decoding Experiment ····· 33

2.4.2 Caption Decoding Experiment ····· 34

2.4.3 Ablation Experiment ····· 40

2.5 Summary and Discussion ····· 41

3. A cross-subject universal decoding method for data migration ····· 44

3.1 Introduction ····· 44

3.2 Experiment data ····· 46

3.3 Model ····· 47
3.4 Experiments in Increasing Subjects Data ····· 49
3.4.1 Experimental Design ····· 49
3.4.2 Results ····· 52
3.5 Subject Data Migration Experiment ····· 55
3.5.1 Experimental Design ····· 55
3.5.2 Results ····· 56
3.6 Equilibrium Parameter Experiment ····· 58
3.6.1 Experimental Design ····· 58
3.6.2 Results ····· 60
3.7 Summary and Discussion ····· 61
4. Conclusion and Discussion ····· 64
4.1 Conclusion ····· 64
4.2 Discussion ····· 66
References ····· 68

Abstract

Neural decoding utilizes machine learning techniques to predict external stimuli or cognitive tasks from brain neural activity, which not only aids in understanding the encoding mechanisms of information in the brain but also finds wide applications in the research and practice of brain-machine interface models. In previous neural decoding studies, due to the structural and functional differences among individuals' brains and the diverse requirements of different decoding tasks, neural decoding models exhibited subject-specific and task-specific characteristics. Establishing a universal decoding model would help reduce the complexity of decoding systems and save training costs. Previous research on universal decoding models has often focused only on either cross-subject or multi-task universality and has not utilized data transfer between subjects to alleviate model overfitting issues. Therefore, this thesis expands the ability of universal decoding from these two aspects. My research is mainly divided into the following two parts:

1. Current universal decoding models often only achieve either cross-subject decoding or multi-task decoding, lacking the ability to simultaneously achieve both types of universality. Therefore, this thesis proposes a multimodal language model trained based on the prompt-tune strategy. This model improves upon the GPT model architecture, enabling on-demand decoding of brain signals for corresponding subjects and tasks under prompt instructions. Experimental results demonstrate that the decoding model successfully achieves universal decoding for multiple subjects and tasks simultaneously, with decoding performance on each subject's data far exceeding chance levels in classification tasks and text description tasks. Meanwhile, the introduced network encoder block further enhances the performance of the decoding model on all decoding tasks, making it perform at the state-of-the-art level in text description tasks.

2. Increasing the number of subjects will increase the retraining cost for universal decoding models, and the large data requirements for decoding models may lead to overfitting issues. This thesis proposes a universal decoding method for cross-subject data transfer to address these issues. This model is based on a scalable autoencoder framework, capable of achieving decupled time and space cross-subject data alignment. The experimental results show that, compared with the traditional functional alignment methods, using the proposed method for cross-subject alignment can effectively improve the performance of universal decoding models on new subject data. By implementing cross-subject data transfer, this method successfully augments the training set with data from other subjects, effectively alleviating

overfitting issues caused by insufficient data from specific subjects. The use of generalized contrastive learning constraints further reduces the demand for data collection.

Abrégé

La décodage neural utilise des techniques d'apprentissage automatique pour prédire les stimuli externes ou les sensations cognitives à partir de l'activité neuronale cérébrale, ce qui aide non seulement à comprendre les mécanismes d'encodage de l'information dans le cerveau, mais trouve également de nombreuses applications dans la recherche et la pratique des modèles d'interface cerveau-machine. Dans les études précédentes sur le décodage neural, en raison des différences structurelles et fonctionnelles entre les cerveaux des individus et des exigences diverses des tâches de décodage, les modèles de décodage neural présentent des caractéristiques spécifiques au sujet et à la tâche. L'établissement d'un modèle de décodage universel aiderait à réduire la complexité des systèmes de décodage et à économiser les coûts de formation. Les recherches précédentes sur les modèles de décodage universels se sont souvent concentrées uniquement sur la universalité inter-sujets ou multi-tâches, sans utiliser le transfert de données entre les sujets pour atténuer les problèmes de surajustement du modèle. Par conséquent, cette thèse élargit la capacité de décodage universel selon ces deux aspects. Ma recherche est principalement divisée en deux parties :

1. Les modèles universels de décodage actuels n'atteignent souvent que le décodage inter-sujets ou multi-tâches, manquant de la capacité à réaliser simultanément les deux types d'universalité. Ainsi, cette thèse propose un modèle de langage multimodal entraîné sur la stratégie de réglage par invite. Ce modèle améliore l'architecture du modèle GPT, permettant le décodage à la demande des signaux cérébraux pour les sujets et tâches correspondants sous des instructions d'invite. Les résultats expérimentaux montrent que le modèle de décodage parvient avec succès à réaliser un décodage universel pour plusieurs sujets et tâches simultanément, avec des performances de décodage sur les données de chaque sujet dépassant largement les niveaux de chance dans les tâches de classification et de description textuelle. Parallèlement, le bloc d'encodeur réseau introduit améliore encore les performances du modèle de décodage sur toutes les tâches de décodage, le faisant fonctionner au niveau de l'état de l'art dans les tâches de description textuelle.

2. L'augmentation du nombre de sujets entraînera des coûts de ré-entrainement pour les modèles de décodage universels, et les grandes exigences en données pour les modèles de décodage peuvent mener à des problèmes de surajustement. Cette thèse propose une méthode de décodage universel pour le transfert de données inter-sujets afin de résoudre ces problèmes. Ce modèle est basé sur un cadre d'auto-encodeur évolutif, capable d'obtenir un alignement des

données inter-sujets découplé en temps et en espace. Les résultats expérimentaux montrent qu'en comparaison avec la méthode traditionnelle d'alignement fonctionnel, l'utilisation de la méthode proposée pour l'alignement inter-sujets peut améliorer efficacement les performances des modèles de décodage universels sur les nouvelles données des sujets. En mettant en œuvre le transfert de données inter-sujets, cette méthode augmente avec succès l'ensemble de données d'entraînement avec des données provenant d'autres sujets, atténuant ainsi efficacement les problèmes de surajustement causés par des données insuffisantes provenant de sujets spécifiques, et l'utilisation de contraintes d'apprentissage contrastif généralisé réduit encore la demande de collecte de données.

Acknowledgements

I wish to convey my heartfelt thanks to Dr. Amir Shmuel and Dr. Huafu Chen for their exceptional mentorship and steadfast support throughout my research. Their guidance has been instrumental in my progress.

I am equally grateful to the McConnell-Brain Imaging Centre and the Brain Imaging and Pattern Recognition Lab, along with their dedicated members, for their collaborative efforts and valuable contributions. I also want to acknowledge Dr. Icaro Oliveira and Dr. Wietske van der Zwaag for their generous provision of data, which has been crucial to my research.

A special note of appreciation is due to my friends and family for their unwavering support and patience during this challenging yet gratifying journey. Their encouragement has been a pillar of strength.

Thank you all for playing such a vital role in my academic and personal development.

List of Figures

Figure 2-1 Schematic diagram of decodeGPT decoding process ····· 22
Figure 2-2 GPT-based universal decoding model architecture ····· 24
Figure 2-3 Classification accuracy of the test set at different stages of training ····· 33
Figure 2-4 Confusion matrix for category-aware experiments ····· 34
Figure 2-5 Partial results of the caption decoding task for four subjects (a) ····· 36
Figure 2-5 Partial results of the caption decoding task for four subjects (b) ····· 37
Figure 2-6 Comparison of the distribution of descriptive text and random level fidelity metrics predicted by the model ····· 39
Figure 3-1. Alignment model for the i-th subject ····· 48
Figure 3-2 Comparison of processes when adding new subject data ····· 50
Figure 3-3 Experimental flow of the added subject data for measuring the degree of public space deviation ····· 51
Figure 3-4 Comparison of the performance of our method with Hyper-Alignment (HA) and Regularized Hyper-Alignment (RHA) when processing new subject data on the ds000105 dataset ····· 53
Figure 3-5 Comparison of the performance of our method with Hyper-Alignment (HA) and Regularized Hyper-Alignment (RHA) when processing new subject data on the ds000117 dataset ····· 54
Figure 3-6 Comparison of the performance of our method with Hyper-Alignment (HA) and Regularized Hyper-Alignment (RHA) when processing new subject data on the BOLD and VASO dataset ····· 55
Figure 3-7 Comparison of the demand for new subject data for the three different data processing methods ····· 57
Figure 3-8 Balancing parameters in the loss function λ effect on alignment performance 61

List of Tables

Table 2-1 Visual Area Information Table ····· 31

Table 2-2 Comparison results of DecodeGPT with several other state-of-the-art caption decoding models on the caption decoding task (bolded numbers represent the highest performance achieved in each metric compared to the other compared models) ····· 40

Table 2-3 Comparison of results in the category decoding task for models that include the network encoder block and those that do not include it (bold numbers indicate the best performance achieved in each metric in the models compared) ····· 40

Table 2-4 Comparison of results in the caption decoding task between models that include the network encoder block and those that do not include it (bold numbers indicate the best performance achieved in each metric in the models compared) ····· 41

List of Acronyms

fMRI	functional Magnetic Resonance Imaging
ROI	Region of Interest
LR	Linear Regression
SVM	Support Vector Machine
CNN	Convolutional Neural Network
RBF	Radial Basis Function
LSTM	Long Short-Term Memory network
VAE	Variational Autoencoder
GAN	Generative Adversarial Network
LDM	Latent Diffusion Models
CLIP	Contrastive Language-Image Pre-training
GUSE	Google Universal Sentence Encoder
GPT	Generative Pre-trained Transformer
AC	Anterior Commissure
PC	Posterior Commissure
HA	HyperAlignment
CCA	Canonical Correlation Analysis
RHA	Regularized HyperAlignment
KHA	Kernel HyperAlignment
DHA	Deep HyperAlignment
DGCCA	Deep Generalized Canonical Correlation Analysis
SHA	Supervised HyperAlignment
GBDM	Graph-Based Decoding Model
RLHF	Reinforcement Learning from Human Feedback
GLM	General Language Model
NSD	Natural Scenes Dataset
BLEU	Bilingual Evaluation Understudy
ROUGE	Recall-Oriented Under-study for Gisting Evaluation
METEOR	Metric for Evaluation of Translation with Explicit Ordering
VT	Ventral Temporal
BSC	Between-Subject Classification

LOO	Leave One Out
BOLD	Blood Oxygen Level Dependent
VASO	Vascular Space Occupancy
BCI	Brain-Computer Interface

1.Introduction

1.1 Statement of Contribution

We extend our gratitude to Dr. Icaro Oliveira and Dr. Wietske van der Zwaag for providing the non-public CBV and BOLD weighted functional dataset. My supervisor, Dr. Amir Shmuel, supported me through every milestone of my academic journey, offering insightful guidance and concrete advice, even while I was working remotely. Additionally, my supervisor, Dr. Huafu Chen, frequently encouraged and motivated me with valuable suggestions. I could not have completed this thesis without his support.

The author's (Diwei Wu) contributions are as follows: Review and summarize the background knowledge in Chapter 1; Propose and implement the universal decoding model, perform data pre-processing on public datasets, and conduct all experiments and analyses in Chapters 2 and 3.

1.2 Neural Decoding

The brain is the center of the nervous system, demonstrating extraordinary complexity and impressive computational power. It dominates important cognitive functions such as consciousness, perception, and higher-level pursuits involving cognitive, emotional, and motor management. Because of its extreme complexity and power, the brain remains the primary focus of extensive research that seeks to demystify its intricate inner workings and solve the mysteries associated with its function.

Neuro decoding is one of the core tools of neuro engineering and neural data analysis, providing us with the means to explain how representations are encoded in the brain. Neuro decoding uses neural activity recorded from the brain to predict stimulus variables in the external world. For example, neural activity in the primary motor cortex is used to predict finger movements (Shin, Aggarwal, Acharya, Schieber, & Thakor, 2010). Neural activity in the language cortex, including the ventral sensorimotor cortex, superior temporal sulcus gyrus, and inferior frontal gyrus, is used to predict speech (Anumanchipalli, Chartier, & Chang, 2019) and neural activity in the visual cortex is used to reconstruct images of visual stimuli (Miyawaki et al., 2008). These studies allow us to understand in which different brain regions the properties of sensory input or motor output are encoded. In addition, neural decoding even allows us to reconstruct internal representations that can only be observed during cognitive processes, such as imagery and dreams (T. Horikawa, Tamaki, Miyawaki, & Kamitani, 2013).

This allows us to quantitatively discover mental and intellectual activities that can only be described qualitatively.

Neural decoding also has important applications in engineering, the most representative application being Brain-Computer Interface (BCI), where neural signals from the motor cortex of the brain are usually captured and the corresponding decoded predictions are used to control an external device, such as a cursor or a robotic arm (Collinger et al., 2013; Serruya, Hatsopoulos, Paninski, Fellows, & Donoghue, 2002). Neurological disorders that result in loss of communication can have a serious impact on the quality of life of patients, and BCIs will help such patients to establish alternative communication devices to regain their ability to communicate through residual non-verbal activities such as brain signals.

We detect neural activity through methods of acquisition of neural signals and study how the brain implements cognitive processes and generates behavior. There are many techniques - electrical, optical, and chemical - that allow us to observe neural activity on different temporal and spatial scales. If we are interested in the activity of a single neuron, we can use electrodes located within the neural tissue to record single or multi-unit activity. This activity reflects action potentials from one or more neurons in the brain and can be used, for example, to determine how neuronal firing rates vary in conjunction with behavioral variables. In addition, we can examine the oscillatory dynamics of neural activity, which can be obtained from various forms of signals such as Local Field Potential (LFP), Electrocorticogram (ECoG), or Electroencephalogram (EEG). The acquisition of signals in various forms, such as electrophysiological recording techniques, functional brain imaging techniques such as functional Magnetic Resonance Imaging (fMRI), and functional Near-Infrared Spectroscopy (fNIRS), provide different perspectives on brain function. Rather than studying electrical signals, they capture the hemodynamic consequences of underlying neural activity.

In practice, neural decoding is usually regarded as a machine learning problem of regressing high-dimensional neural signals on high-dimensional stimulus variables. Therefore, depending on the way the neural signals are acquired, the recording point location, or the Region of Interest (ROI), different decoding methods have been applied in past studies. Linear Regression (LR), Support Vector Machine (SVM), and Neural Networks are the most widely used methods. Different decoding methods usually make different implicit assumptions about the data; for example, Regularized Linear Regression is based on the assumption that the output varies proportionally to the input, in which any noise contained in the output is treated as

Gaussian noise. Decoders often make assumptions about the mapping of inputs and outputs; some methods, including linear regression and Kalman filtering, assume that the mapping between inputs and outputs is linear, whereas others, including neural networks, assume that there is a flexible non-linear mapping between inputs and outputs. Models based on linear assumptions tend to perform better when the amount of data is small, or there is a lot of noise, but when the relationship between the inputs and outputs is more agnostic, models using nonlinear assumptions are superior solutions. In addition to this, different classes of machine learning methods will provide different forms of estimates for decoding. Conventional machine algorithms will provide maximum likelihood estimates of the decoded variables, i.e., single-point estimates that are most likely to be true. Bayesian decoding, on the other hand, produces posterior distributions as decoded outputs and can, therefore, provide information about the uncertainty of the estimates. Mathematically, the maximum likelihood estimate is the vertex value of the posterior distribution under the condition that the prior distribution is uniform.

1.3 Visual Neural Decoding Task

The brain exhibits dynamic responses to visual stimuli received through the eyes and shows unique response patterns when exposed to various visual stimuli (Tomoyasu Horikawa & Kamitani, 2017; Teng & Kravitz, 2019). Different visual stimuli, such as color, shape, motion, and objects, also trigger specific patterns of neural activity within the brain.

Visual decoding research involves several directions, each of which is dedicated to understanding and restoring different aspects of the brain's perceptual content of visual stimuli. The three main directions represented in the current research are category decoding, caption decoding, and reconstruction decoding. Among them, category decoding focuses on identifying and classifying the main categories of visual stimuli in brain activity, aiming to reveal how the brain recognizes and understands different types of visual information (Tomoyasu Horikawa & Kamitani, 2017; W. Huang et al., 2020; Kaiser, Azzalini, & Peelen, 2016). The goal of category decoding is to predict one or more semantic labels appearing in a visual stimulus by analyzing brain activity and trying to understand how the brain associates semantic concepts with visual information (Huth, Nishimoto, Vu, & Gallant, 2012; Nishida & Nishimoto, 2018). Caption decoding aims at generating textual descriptions of visual stimuli from brain activity. Textual descriptions provide a more detailed description of the scene and the relationship between objects in the scene than simple category labels. As a bridge between visual signals and human language, caption decoding can help us understand how the brain associates visual

and verbal information (Huang, Yan, Cheng, Wang, Li, et al., 2021; Huang, Yan, Cheng, Wang, Wang, et al., 2021; Takada, Togo, Ogawa, & Haseyama, 2020). The task of reconstructive decoding, on the other hand, is to reconstruct the visual information from the visual signal. While the task of reconstructive decoding is to restore the original visual stimulus from brain activity at the pixel level, aiming to explore how the brain encodes and stores visual information (Wei Huang et al., 2020; Huang, Yan, Wang, et al., 2021; Rakhimberdina, Jodelet, Liu, & Murata, 2021). Together, these decoding directions constitute a multilevel study of the understanding of visual neural activity, providing important insights into our deeper understanding of how the brain processes and interprets visual information.

1.3.1 Category Decoding

Early work on visual decoding focused exclusively on identifying object category labels in visual stimuli and advanced the field of computational neuroscience. Haxby et al. used SVM classifiers to classify visual stimulus-response patterns in the ventral temporal cortex and successfully identified grey-scale image stimuli for eight classes of objects(J. V. Haxby et al., 2001) . To address the performance limitations of linear SVM classifiers, Song et al. used a nonlinear radial basis function (RBF) kernel SVM to achieve higher decoding accuracy (Song, Zhan, Long, Zhang, & Yao, 2011). Further, considering a certain delay due to brain hemodynamics, Huang et al. used a deep decoding model based on LSTM to decode categories of brain response sequences for a period of time after receiving a stimulus and achieved a better performance than single time point decoding.

1.3.2 Reconstruct Decoding

Then, fMRI-based brain decoding techniques have evolved from basic fMRI classification methods to more sophisticated methods such as image reconstruction from fMRI (Miyawaki et al., 2008; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009; Nishimoto et al., 2011). This progress has extended the capabilities of fMRI analysis, allowing us to decode and reconstruct visual information from patterns of brain activity. Deep generative models, including variational autoencoders (VAEs), generative adversarial networks (GANs), and latent diffusion models (LDMs), have revolutionized visual reconstruction techniques. These models have been widely used to reconstruct complete images by mapping brain signals to latent variables (S. Lin, Sprague, & Singh, 2022; Shen, Horikawa, Majima, & Kamitani, 2019; VanRullen & Reddy, 2019). Lin et al. introduced DCNN-GAN, a model that combines a reconstruction network and a GAN module. The model employs CNNs for hierarchical feature extraction and

uses DCNN-GAN to transform more realistic images from fMRI signals to pixel space through the reconstruction process (Y. Lin, Li, & Wang, 2019). Fang et al. proposed Shape Semantic GAN to retain semantic information in visual stimulus images. It takes into account the functional differences in visual cortical regions and uses shape and semantic guidance for reconstruction (Fang, Qi, & Pan, 2020).

Although these studies are able to capture forms, colors or images similar to the original stimuli, a common problem observed is that reconstructions are often ambiguous and may contain mixed elements from unrelated concepts. The emergence of diffusion models can help with this problem. Takagi et al. used a stable diffusion LDM called Stable Diffusion, which was built on top of LDMs and trained to reconstruct visual stimuli from very large datasets (Takagi & Nishimoto, 2023). By accepting control from latent representations from the higher cortex, images with high semantic fidelity were generated.

1.3.3 Caption Decoding

In addition to classification and reconstruction, another way we can understand the neural correlates of visual perception is through fMRI text description generation. By converting fMRI signals into human-understandable language, we can focus on "high-level" attributes (e.g., object categories) and "higher-level" attributes (e.g., semantic scene descriptions) rather than "low-level" attributes (e.g., oriented edges) and "mid-level" attributes (e.g., texture). There have been research attempts to estimate semantic information by correlating brain activity signals with words or sentences representing visual stimuli. Initially, fMRI text description generation was based on image caption generation techniques. Matsuo et al. proposed a method for generating captions for visually perceived images based on image caption generation models using fMRI data (Matsuo, Kobayashi, Nishimoto, Nishida, & Asoh, 2018). In this method, captions are generated by converting fMRI data first to image features and then to text features. However, there is a potential risk of losing important information in the fMRI data during these two stages of conversion. Therefore, Takada et al. proposed a method to generate subtitles directly from fMRI data via unsupervised text latent space (Takada et al., 2020). Huang et al. used a language model as a framework to generate caption decoding content directly from fMRI neural activity stimulated by natural images (Huang, Yan, Cheng, Wang, Li, et al., 2021).

Recently, increasing decoding efforts have begun to use CLIP or pre-trained language models to help learn the relationship between brain signals and images or text in order to

reconstruct images or generate captions based on brain signals. By pre-training on a large amount of material, CLIP can master complex mapping relationships between images and text, such as recognizing objects, and understanding attributes and concepts conveyed through language. Understanding and exploiting the associations between images and text is crucial in tasks involving multiple modalities, and this is where CLIP excels (Radford et al., 2021).

These methods typically rely on a latent feature space learned by CLIP or pre-trained language models that encode rich semantic information extracted from large amounts of text or images. They first map brain signals to this semantic space and then use downstream caption decoders to generate corresponding text descriptions. For example, Doerig et al. mapped cortical activity to the GUSE (Google Universal Sentence Encoder) semantic embedding vector space and found the sentences that were closest to the predicted (Doerig et al., 2022). Chatterjee and Samanta proposed Dreamcatcher, which encodes brain fMRI signals into a text embedding space based on a pre-trained large-scale language model, GPT, and then uses the downstream language model for text generation (Pina et al., 2020). The work of Ferrante et al. is based on the GIT (Generative Image-to-text Transformer) framework, in which they encode fMRI signals into the text-image representation space of a pre-trained CLIP model, and then generate text descriptions based on the corresponding pre-trained GPT decoder (Ferrante, Ozcelik, Boccato, VanRullen, & Toschi, 2023). Luo et al. projected the weights into the CLIP embedding space of natural images and generated sentences by decoding the voxel-by-voxel weights (Luo, Henderson, Tarr, & Wehbe, 2023).

1.4 Cross-subject Data Alignment

The problem of individual differences is an important challenge that cannot be ignored in neuroimaging research. Significant differences exist in anatomical structures and functional patterns between individuals, and such differences will affect the ability to interpret, analyze, and generalize data from multiple subjects. Individual variability can be due to a variety of genetic, developmental, environmental, and lifestyle factors, and the main areas of variability include but are not limited to the thickness of the cerebral cortex, the distribution of grey and white matter, and the size and shape of individual brain regions. At the same time, the functional activity patterns of the brain may also vary greatly from one individual to another due to structural differences, cognitive characteristics, and learning history. That is to say, under exactly the same task or stimulus conditions, the brain activity patterns of different individuals may be very different.

1.4.1 Anatomical Alignment

Currently, anatomical alignment based on structural images is widely used in multi-subject fMRI analysis tasks. Anatomical alignment is a process of aligning brain structures from different individuals into a standard anatomical space and is often used for the preprocessing of fMRI data. The alignment algorithms involved in anatomical alignment mainly include linear alignment, nonlinear alignment, and local alignment, and the purpose of the alignment is to make the same location in the data of different subjects have the same position, orientation, and scale in the same space. Specifically, anatomical alignment acquires image features for alignment from the original image, such as edges and corner points. The extracted features are then matched with corresponding features in the target space to determine the correspondence between the original data and the target space and to select a suitable transformation model, such as a rigid transformation including translation, rotation, and scaling, an affine transformation including translation, rotation and scaling while preserving parallelism, and a non-rigid transformation including elastic deformation. After optimization with the objective of minimizing the distance between matched feature points, the optimal transformation parameters are obtained, and the original image data are transformed accordingly. The transformation process includes steps such as spatial coordinate conversion, pixel interpolation, and resampling. In order to obtain a unified reference frame for the human brain, various standardized brain spaces have been proposed.

The Talairach space is a standardized space based on brain anatomy developed by neuroscientist Jean Talairach. It is based on a large amount of brain anatomical data and describes the structure and function of the brain by converting brain structures into a three-dimensional coordinate system. The coordinate system is the AC-PC coordinate system, i.e., a spatial coordinate system based on the Anterior Commissure (AC) and Posterior Commissure (PC). The Talairach Standard Brain Atlas is also based on a large amount of brain anatomical data describing the different structures and regions of the brain and is, therefore, of great use in Functional brain imaging studies related to functional localization and regional analysis. The MNI (Montreal Neurological Institute) space is another brain standardized space developed by the Montreal Neuroscience Institute in Canada, which has a coordinate system based on the central axes of the skull, orbits, and ear holes, with the origin of the coordinates located at the head's central. The origin of the coordinates is located at the center of the head. For the Talairach coordinate system, the three dimensions represent the offset of the voxel point

relative to the AC in the anterior-posterior, up-down, and left-right directions, whereas in the MNI coordinate system, the three dimensions represent the offset relative to the origin of the MNI coordinate system. Therefore, although a coordinate system more directly related to anatomical structures may have been more applicable in early functional brain imaging studies, MNI space has been more widely used in MRI data analysis for its more intuitive and easy-to-understand advantages. Currently, the standard brain atlas commonly used in MNI space is MNI152, a standard brain structure atlas created based on MRI data from 152 individual subjects, which has a higher spatial resolution and a more accurate delineation of the brain structure compared to Talairach's atlas.

1.4.2 Functional Alignment

However, the accuracy of these alignment methods is still limited by the differences in the size, shape, and anatomical location of functional regions between subjects. There are often large differences in the anatomical structure of the brain between individuals, such as differences in the shape, size, and location of the sulcus gyrus, which can be difficult to completely eliminate even with high-level alignment algorithms, leading to errors in the alignment. In addition, the setting of parameters and the selection of markers during the alignment process can also affect the accuracy of anatomical alignment.

To overcome the limitations of anatomical alignment, functional alignment was proposed. Functional alignment is based on functional imaging and utilizes multi-view learning methods to achieve better alignment results than anatomical alignment. The first functional alignment method, HyperAlignment (HA), proposed by Haxby et al., laid the foundation for subsequent functional alignment methods (James V Haxby et al., 2011) . HyperAlignment is based on the assumption that cortical response pattern vectors in the brains of two individuals receiving the same stimulus (e.g., watching the same full-action film) reflect similar information, however, their coordinate systems representing their respective spaces are not aligned. HyperAlignment uses Procrustes transformations to iteratively process pairs of test samples to derive a population coordinate system in which the vector trajectories of each pair of test subjects are in the best possible alignment after optimization. HyperAlignment can be regarded as a multi-view learning method. It is mathematically related to Canonical Correlation Analysis (CCA) and essentially differs only in its constraints. After realizing this, Xu et al. proposed Regularized HyperAlignment (RHA) (Xu, Lorbert, Ramadge, Guntupalli, & Haxby, 2012). By introducing a regularization method, Regularized HyperAlignment achieves better

performance in alignment, even when compared with various subsequent methods, and can achieve relatively high levels of performance (Li, Liu, Chen, & Zhang, 2020). With the development of multi-view learning methods, many other variants of functional alignment have arisen.

Lorbert et al. proposed Kernel HyperAlignment (KHA) based on the Kernel Canonical Correlation Analysis (CCA) method, which solves common nonlinear and high-dimensional problems in the representation space (Akaho, 2006). This method solves nonlinear and high-dimensional problems in common representation spaces (Lorbert & Ramadge, 2012). Chen et al. developed a new alignment model (Shared Response Model, SRM) based on the Shared Response Assumption, which implicitly learns shared patterns across subjects and can be considered as a probabilistic CCA (P.-H. C. Chen et al., 2015) . With the development of deep learning, CCA has been endowed with more powerful tools and methods, and significant progress has been made in dealing with large-scale data and nonlinear relationships. Yousefnezhad et al. proposed the Deep HyperAlignment (DHA) method by taking advantage of the power of deep neural networks, which uses deep neural networks as the kernel function as a kernel function, and in doing so, eliminates the performance limitations of fixed kernel functions (Yousefnezhad & Zhang, 2017). DHA uses the deep neural network as the kernel function and eliminates the performance limitation of the fixed kernel function to achieve excellent alignment performance. Deep HyperAlignment can also be seen as a variant of Deep Generalized Canonical Correlation Analysis (DGCCA) (Benton et al., 2017). To further improve the alignment performance, they also introduced Supervised HyperAlignment (SHA), which achieves even better alignment performance by introducing additional label information for supervised learning (Yousefnezhad, Selvitella, Han, Zhang, & Systems, 2020). Various approaches based on deep neural networks have achieved excellent results in improving functional alignment performance. Another aspect of functional alignment lies in exploring the practicality of functional alignment. For example, existing functional alignment methods are based on temporally aligned fMRI data, i.e., brain signals acquired by different subjects receiving the same stimulus sequences synchronously. In order to move away from this to a certain extent, Li et al. proposed the Graph-Based Decoding Model (GBDM), which makes use of cross-subject attempts to characterize the similarities and differences between all samples, thus allowing the method to deal with fMRI data not temporally aligned (Li et al., 2020).

1.5 GPT-based Large Language Model

GPT is a language model proposed by OPENAI in 2018, which means "Generative Pre-Trained Transformer" (Radford, Narasimhan, Salimans, & Sutskever, 2018). "Generative" means that GPT is a language model for text generation tasks, and GPT adopts a unidirectional Transformer architecture in the model structure, i.e., only the decoder part of the Transformer is used (Vaswani et al., 2017). GPT, therefore, does not rely on the contextual information provided by the encoder and autoregressively makes predictions of what follows using only the textual information from above on the left side. In a language generation task, the generator needs to predict what comes next based on what has already been generated and should not rely on future information. Therefore, the unidirectional contextual information of GPT is more in line with the needs of generative tasks and thus performs well on generative tasks. On the other hand, "pre-training" represents the unsupervised pre-training strategy in the training phase of GPT. The first phase of GPT training is unsupervised pre-training using a massive text corpus from the Internet, in which the model learns a wide range of linguistic knowledge and semantic representations that are not task- or domain-specific and can be further trained in the subsequent training phase. Training phases GPT can be further fine-tuned for specific tasks to further enhance its performance in particular tasks.

GPT-2 uses a larger number of parameters and a deeper network structure compared to GPT and expands the training data, allowing the model to better understand linguistic information and produce higher quality text, and the unsupervised pre-training phase on a much larger amount of data gives GPT the ability to generalize to tasks or domains that have never been seen before (Radford et al., 2019). GPT-3 is a new generation of GPT. GPT-3 has even improved from 1.5 billion parameters in GPT-2 to 175 billion, and the amount of pre-training data has been increased from 40 GB to 45 TB, which allows GPT-3 to outperform zero-shot or few-shot SOTA methods on most tasks without relying on fine-tuning for a specific task at all, and even on tasks such as mathematical addition and writing code (Brown et al., 2020). GPT-3.5 introduces prompt-tune and Reinforcement Learning from Human Feedback (RLHF) on top of GPT-3, which enables GPT-3.5 to generate high-quality content that is more in line with human preferences. By aligning the large model with human preferences, GPT-3.5 is able to generate high-quality content that is more in line with human preferences. By aligning large models with human preferences, GPT-3.5 demonstrates great intelligence, and the powerful intent understanding and response generation capability of ChatGPT based on GPT-3.5 is good

proof of this great breakthrough (Ouyang et al., 2022). GPT-4, on the other hand, extends the capability of the large language model to multimodality, enabling the model to understand both human language and image inputs at the same time, elevating the understanding capability of the large language model to a new dimension (Achiam et al., 2023).

Due to the great success of the GPT series of models proposed by OPENAI, other big language models based on the GPT style have emerged. For example, the PaLM series of large language models released by Google (Anil et al., 2023; Chowdhery et al., 2023) and Meta's LLaMA series of models (Touvron, Lavril, et al., 2023; Touvron, Martin, et al., 2023). Tsinghua University also proposed a non-GPT-style language model GLM, i.e., a General language model, and trained a large language model ChatGLM based on it. GLM is also based on the Transformer architecture but differs from GPT in that GLM takes the autoregressive fill-in-the-blanks as the training goal instead of the GPT's autoregressive style following prediction (Z. Du et al., 2021). Moreover, GLM employs a bidirectional attention mechanism, which can theoretically understand the overall context better than GPT-style models.

Based on the above types of base GLMs trained on massive predictive data, various types of stylized GLMs with better performance in niche areas through fine-tuning have also demonstrated excellent performance and broad prospects. For example, Bloomberg GPT, which is trained on a large amount of financial data sources, can automatically generate high-quality financial reports with given topics and contexts, as well as refine and sort out financial news and financial information (Wu et al., 2023). In the medical field, the medical model Med-PaLM2 (Singhal et al., 2023) has achieved an accuracy score of 85 on the U.S. Physician's Licensing Exam, reaching the level of an "expert" candidate. In this paper, we will also propose a multimodal GPT language model for promote-tune for brain decoding tasks, so as to build a general visual decoding model that can achieve cross-subject and multi-task decoding.

1.6 Development in Universal Neural Decoding

Whether it is classification decoding, caption decoding or reconstruction decoding, the corresponding work has continuously optimized the model structure for a specific decoding goal, so if we need to perform different decoding tasks, we also need to switch different model structures. In addition to this, due to the variability of brain structure and function between individuals, the response patterns of fMRI signals from each subject are also vastly different, so a decoding model trained on the data of one subject cannot directly process the data of other subjects. This also leads to the fact that we have to re-train the decoding model for a specific

subject and a specific task when facing different subjects and different tasks. This will limit the generality and applicability of the models and significantly increase the development and deployment cost of the decoding models.

Regarding the two problems mentioned above, the generalization of decoding models can be summarized in the following two directions: the multi-task decoding direction and the cross-subject decoding direction. The goal of the multitask decoding direction is to achieve decoding for different decoding tasks with as little complexity as possible within a single model. In addition to improving the generality of the model, multitask decoding will help explore the brain's processing mechanism for common information across different tasks. The goal of cross-subject decoding is to achieve decoding of different subjects' data within a single model with as little complexity as possible. Compared to the number of tasks, there is a large scope for developing the number of individual subjects. Therefore, in addition to the basic need for generality, cross-subject modeling should also focus on the scalability of the model to the growing number of subjects and how to effectively utilize brain signals from different subjects with different content to reduce the cost of data acquisition for a single subject is also a valuable issue.

In the area of visual decoding, some recent work has begun to experiment with more generalized visual decoding.

In the direction of multi-task decoding, Mai proposed the UniBrain architecture, which allows for simultaneous visual stimulus reconstruction and caption decoding tasks in a diffusion model. Nonetheless, in the final stage of decoding using this framework, it is still necessary to specify specific decoders for different decoding tasks for specific forms of content generation. In addition to this, they did not make any attempts in the direction of cross-subject decoding. In fact, they trained separate subject-specific models for each of the four subjects selected from the NSD dataset, which further increased the complexity of the decoding system (Mai & Zhang, 2023).

In the direction of cross-subject decoding, Matteo et al. used a functional alignment technique based on hyper-alignment to address this problem. Although finally they implemented a cross-subject single-task image reconstruction decoding model, additional alignment processing of cross-subject data is still required and additional alignment models need to be trained for the subject data alignment process (Ferrante, Boccato, & Toschi, 2023).

No visual decoding research has yet been able to accomplish general decoding in both directions, multi-task and cross-subject.

1.7 Research Content and Significance

Decoding of fMRI neural signals is limited by inter-subject variability as well as variability in task requirements, thus requiring the training of specific decoding models for each subject and each task, which increases the cost of training and deploying decoding models. There have been studies that have attempted to enhance the generality of decoding in one of these areas. However, there are no studies that have addressed both cross-subject and cross-task decoding. Meanwhile, due to the scalability of the number of subjects, the growing number of subjects in a cross-subject universal decoding model will also bring a series of problems. Therefore, in this paper, we construct a universal visual decoding model based on GPT, which can achieve cross-subject and multi-task decoding within a single model according to the Prompt instruction. We also propose a cross-subject data migration method, which can solve the retraining problem of the universal model brought about by the ever-growing number of subject data and reduce the amount of training data required for the universal model effectively through inter-subject data migration.

The research covers two directions, as summarized. The first study considers cross-subject multitasks universal decoding based on GPT and Prompt techniques. The current decoding research in the direction of enhancing the generality of decoding models is still limited to only one aspect of cross-subject generality and multi-task generality, and no research has been done to achieve both generalities simultaneously. Therefore, our study firstly establishes a cross-subject and multi-task generalized visual decoding model based on the GPT model and Prompt technology, and successfully realizes the decoding of multiple subjects' data and multiple tasks within a single model. The experimental results show that the proposed model performs significantly better than the chance level on the visual stimulus categorization task, and significantly better than the other caption decoding methods on the caption decoding task. We also introduce a whole-brain information interaction module to enhance the generality of decoding model. The introduced whole-brain information interaction module further improves the performance of classification decoding and caption decoding. The GPT-based universal decoding model architecture proposed in this study will help to save the training and testing overheads of building subject-specific and task-specific models, and the use of GPT pre-trained on a large corpus as a text generator will help to improve the performance of caption decoding.

Universal decoding research oriented towards subject data migration. In the direction of cross-subject universal decoding, there have been studies using functional alignment methods to reduce the variability of data between subjects and build cross-subject universal decoding models. However, due to the scalability of the number of subjects, it is necessary to re-do the cross-subject alignment and re-train the universal decoding model after adding data from new subjects, and the time overhead of the alignment process will grow exponentially with the growing number of individual subjects. Therefore, our second study establishes a scalable feature alignment method, which can align the new subjects' data to the original common feature space so that the original universal decoding model can process the new subjects' data without re-training. The alignment process of the new subjects' data is decoupled from all the previous subjects' data in time and space, which achieves the linear growth of the time overhead in the computation process. Meanwhile, the method established in this paper achieves effective subject data migration, which effectively reduces the demand of the decoding model for a specific amount of subject data by expanding the number of training sets brought by migrating other subject data.

2. Multi-task cross-subject universal decoding model based on GPT

2.1 Introduction

Visual decoding in the brain is a research area in neuroscience and computer science, aiming to reveal the brain's mechanisms for processing and understanding visual information. However, although many studies have proposed various decoding models for visual signals, these decoding models face a common challenge of difficulty in achieving generalizability across individual subjects or across tasks (Han et al., 2019; Y. Liu, Ma, Zhou, Zhu, & Zheng, 2023b; Wen et al., 2018). This means that these models need to design the model architecture and retrain the models individually when facing different subjects or different tasks, which limits their flexibility and generalizability in practical applications. How to solve this problem is an important direction for research in the field of visual neural decoding, and the exploration in this direction will promote the further development of visual decoding in the direction of generality.

The emergence of Prompt technology offers a promising solution to this problem. Prompt-Tuning is a technology first developed by GPT-3 (Brown et al., 2020) and PET (Schick & Schütze, 2020). A fine-tuning paradigm is proposed to avoid introducing additional parameters by adding templates, thus enabling language models to achieve desirable results in small-sample or zero-sample scenarios. Prompt is essentially an instruction to a downstream task, which can be viewed as a form of information enhancement. Prompt instructions, when fused with the language model input information, will highlight the corresponding task characteristics, making the language model more likely to produce high quality predictions. If we can design the key information controlling the decoding behavior of the model, such as the "subject + task" information, into the Prompt instruction in the form of natural language, we can achieve multi-task and multi-subject universal decoding with Prompt control on a single model. Due to the breakthrough in text comprehension and generation capability demonstrated by the large language model represented by chatGPT, Prompt-related techniques have also been applied to the field of visual decoding after they were proposed. Sun et al. previously used Prompt-tuned and Fine-tuned language models, respectively, to generate representations of labeled sentences and match the brain through the similarity of representations to signals that were decoded. This type of decoding based on matching existing text can be considered a form of category decoding, and thus, this study has yet to extend the model's capabilities for multitasking or cross-subject decoding through Prompt. At this time, no work has been done to

implement Prompt-controlled commands in a "subject + task" format that would extend the decoder's decoding capabilities.

In addition to this, In the field of visual decoding, Region of Interest (ROI) is usually derived from two sources: the first is from cortical regions identified by current knowledge in neuroscience, and the second is data-driven, i.e., analyzed from the fMRI data itself. The brain's response patterns to natural visual stimuli usually involve a number of large-scale brain networks, such as visual, emotional, attentional, working memory, linguistic, and semantic (Bressler & Menon, 2010). For the first ROI segmentation method, previous studies usually only involve the visual cortex, and completely ignore the possible contribution of signals from other cortices involved in large-scale brain networks. For the second ROI segmentation method, determining the locations and sizes of all functional networks is time-consuming and costly due to the limitations of the task during fMRI experimental design and signal acquisition, and accurately detecting activations from fMRI data remains a challenging problem (M. Chen et al., 2014). Therefore, utilizing global information about the brain has the potential to enhance decoding performance, taking into account the influence of other large-scale networks on the response to visual stimulus. But how to effectively utilize the information contained in all voxels throughout the brain remains an open and challenging question.

Facing the above problems and based on GPT-2 (Radford et al., 2019), a cross-subject and multi-task generalized generative visual decoding model (decodeGPT) is proposed in this paper. On the NSD dataset (Allen et al., 2022), the model combines visual response activities and whole-brain response activities. The model achieves universal decoding of categories and texts through the control instructions of "subject + task". The decoding flow of the model is shown in Figure 2-1. Compared with previous visual decoding models, the model proposed in this paper has two advantages: 1. The "Text-Prompt" controlled decoding model was created, which can simultaneously process different subjects and different tasks under a unified framework and a set of parameters, without the need to build a unique model for each subject and each task, thus achieving universal decoding of cross-subject category and text information; 2. A multi-head cross-attention module was designed, which can capture and utilize the effective information from all cortical regions of the brain through the introduction of whole-brain information and thus improve the decoding performance to a certain extent.

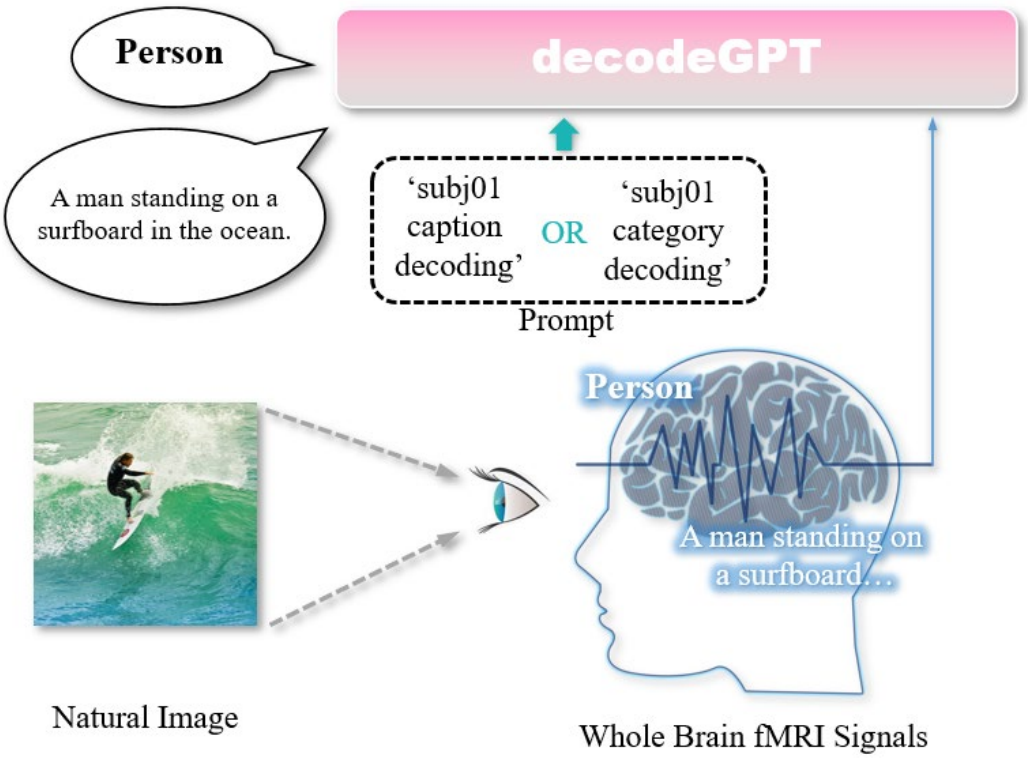

Figure 2-1 Schematic diagram of decodeGPT decoding process

2.2 Experiment Data

In this paper, we use the Natural Scenes Dataset (NSD) to train and test the proposed decoding model (http://naturalscenesdataset.org/). The NSD contains natural image stimuli and corresponding whole-brain responses to natural image stimuli from eight participants. During the nearly one-year data collection period, each participant underwent 30-40 7T fMRI scans using a whole-brain gradient-echo EPI (echo imaging) technique with a pixel size of 1.8 mm isotropic and a repetition time (TR) of 1.6 seconds. During the scan, participants were presented with 9,000-10,000 color images of natural scenes for a total of 22,000 to 30,000 trials. Of the 10,000 images viewed by each participant, 1,000 images were shared by all participants, while the remaining 9,000 images were unique to each individual. The images used in NSD were taken from the Microsoft Common Objects in Context (COCO) dataset (T.-Y. Lin et al., 2014). Each image in the dataset was cropped to a 425 × 425 rectangle. Each image in the COCO dataset is described by five to six natural language sentences provided by humans. During the experiment, each image was displayed for 3 seconds, followed by a 1-second blank interval in which subjects were asked to fixate their eyes on the center of the screen. In addition, each

subject was asked to perform a long-term continuous recognition task to encourage sustained attention to the images.

In the fMRI signal preprocessing stage, the NSD dataset was temporally and spatially interpolated to the fMRI data for slice time correction and head motion correction. Generalized Linear Models (GLM) were then used to estimate the individual trial β weights to represent the subject's voxel-by-voxel response to the presented image stimuli. In the NSD dataset, this paper uses data from four subjects who completed all imaging sessions (subj01, subj02, subj05, and subj07). For each subject this paper uses 27,750 trials out of 30,000 trials, of which 2,250 trials have not been publicly released. Of the 27,750 trials corresponding to samples, 2,770 samples were divided into a test dataset. Included in this test dataset are brain response data after 982 different images were viewed by four subjects, while the remaining trials (N=24,980) were designated as the training dataset.

2.3 Model

In this paper, we propose a multi-task cross-subject visual semantic decoding framework based on the GPT model and Prompt technique. The model can complete the task of semantic decoding or category decoding for fMRI signals from different subjects with different Prompt prompts. The specific decoding process is shown in Figure 2-2.

The decoding model receives fMRI signals from visual areas and other regions. Then, based on the input Prompt cue words, the model will output the corresponding decoding results in the form of natural language sentences. (A) Network encoder: using a multi-head cross-attention architecture, it receives fMRI features from visual regions and fMRI features from all brain regions as input. The fMRI features from all brain regions are used as context for cross-attention computation with the visual region signals to fuse the whole brain information and enhance the visual region signals. Subsequently, the features fused with whole-brain information are encoded into the latent feature space through the full connectivity layer and mean computation. (B) Prompt embedding: subject number and decoding task information are entered in the form of cue text, which is then embedded into the potential feature space. (C) Visual encoder: fMRI signals from visual areas were collected while participants received visual stimuli. They are then transformed to the embedding dimension to form a sequence of ROI signals. The visual encoder performs feature fusion of fMRI features from visual areas via GRU and encodes them into the latent feature space. (D) GPT decoder: adds potential features from the cue embedding block, the visual coder and the network encoder to the embedding of

each Prompt cue word, and inputs the fused sequence of feature embeddings into the GPT-2 decoding model. The GPT-2 decoding model outputs the predicted features. Based on the output features, the next word is predicted until the sentence terminator is predicted.

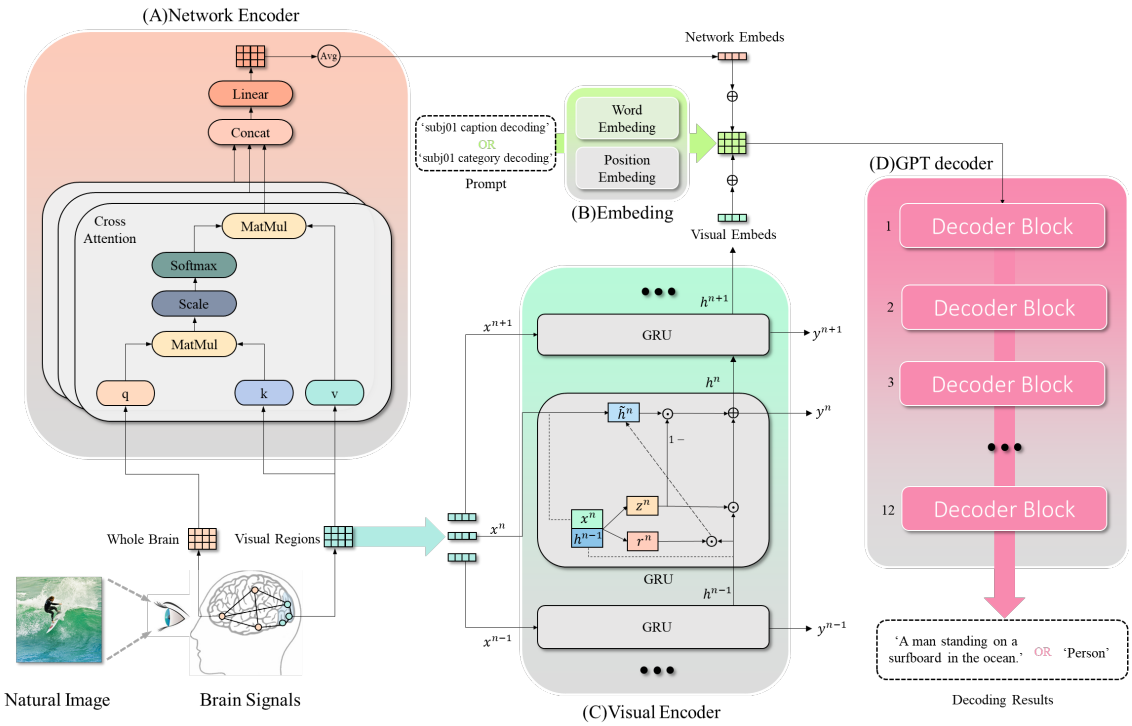

Figure 2-2 GPT-based universal decoding model architecture

2.3.1 Prompt Design

Prompts are task descriptions in textual form that can be used as contextual information to provide guidance for the reasoning process of language modeling. Prompt technology has been widely used in the field of NLP as a method of paradigm transformation between different tasks. In this paper, we adopt Prompt technology to achieve cross-subject universal decoding. Specifically, the Prompt designed in this paper contains the subject number information and the decoding task information. Take "Subject 01 category decoding:" as an example. The "Subject 01" part of the Prompt is the subject number information, which can guide the downstream language model to distinguish the brain signal data from different subjects, and "Subject 01" represents decoding the data from the subject numbered 01. The "category decoding" part is the task information, which will guide the downstream language model to output task-specific decoding results, and "category decoding" represents the category decoding task.

In order to transform the words in a text message into a mathematical representation for model inference, Prompt has to be mapped into a sentence vector by Tokenization before it is

fed into the model T_p; suppose the number of words in the sentence is M, $T_p = (t_1, t_2, \cdots, t_M)$, t_m is the word index of the corresponding word in the glossary.

2.3.2 Word Embedding Module

Word embedding is an effective technique often used in Natural Language Processing (NLP) to represent words as dense vectors that capture their contextual meaning. This is usually done by analyzing word co-occurrence patterns in large text corpora, using techniques such as word2vec (Mikolov, Chen, Corrado, & Dean, 2013), GloVe (Pennington, Socher, & Manning, 2014), and fastText (Bojanowski, Grave, Joulin, & Mikolov, 2017). These models learn vector representations that reflect the semantic relationships between words. By utilizing the information obtained from these embedding models, words with similar meanings or usage patterns are represented as vectors that are closer together in the embedding space. This allows NLP algorithms to understand semantic similarities and relationships between words in a more meaningful way. Word embeddings have proven beneficial in a variety of NLP tasks, including, but not limited to, text categorization, information retrieval, and machine translation. By providing numerical representations of words that encode contextual meaning, word embeddings improve the performance and accuracy of these tasks. Overall, word embeddings play a crucial role in bridging the gap between natural language and machine learning by transforming textual data into a format that algorithms can process and understand efficiently. GPT (Generative Pre-trained Transformer) is a well-known language model developed by OpenAI (Radford et al., 2018). It is structurally based on the Transformer architecture and has attracted much attention in the NLP field due to its performance on generative tasks. Unlike traditional fixed word embeddings, GPT uses contextual word embeddings, also known as contextual representations. Instead of assigning a static vector representation to each word, GPT generates dynamic word representations that consider the context of the target word, its surrounding words, and the entire sentence or document. This contextual word embedding enables GPT to capture the contextually relevant meanings of words and their semantic relationships. By combining the surrounding words and contextual information, GPT's word embedding provides a more nuanced understanding of the language. This allows GPT to produce text that is coherent and appropriate to the context.

In the framework of the proposed model, the embedding part converts the Prompt cue words containing information about the subject and the task into the embedded form and fuses them with other features at the embedding level in the subsequent process. In the output stage,

the model then converts the output of the GPT-2 decoder into the corresponding tokens with maximum probability, which are then converted into text.

The word embedding module embeds word information from text information into the word vector space of the language model, which consists of a mapping matrix $\mathrm{M_{wte}}$ with learnable parameters. Specifically, the word embedding matrix is multiplied by a vector of one-hot forms of each t_m of the Prompt sentence vectors to obtain the representation of each word in the word vector space. The formula for the word embedding module is given in Equation (2-1):

$$\{\mathrm{E_1^w}, \mathrm{E_2^w}, \cdots, \mathrm{E_M^w}\} = \mathrm{onehot}(\mathrm{t_1}, \mathrm{t_2}, \cdots, \mathrm{t_M}) \times \mathrm{M_{wte}} \quad (2\text{-}1)$$

Among them, $\mathrm{E_m^w}$ is the the word vector corresponding to the $\mathrm{t_m}$ word. The one-hot function transforms the numbers into one-hot vectors of the length of the word list. The embedding matrix $\mathrm{M_{wte}}$'s parameters are optimized during the training process so that semantically similar words have similar distances in the word vector space.

2.3.3 Position Embedding Module

The word embedding module simply transforms the information about the words in a sentence into a mathematical representation but does not include information about the position of the words in the sentence. The positional embedding module embeds the positional information of the words in the textual information into the feature space of the language model, which consists of a parameter learnable mapping matrix $\mathrm{M_{wpe}}$. Specifically, the positional embedding matrix is multiplied by a vector $\mathrm{T_p}$ in the Prompt sentence of one-hot forms to obtain a representation of each position in the sentence in the position vector space. The formula for the position embedding module is given in Equation (2-2):

$$\left\{\mathrm{E_1^p}, \mathrm{E_2^p}, \cdots, \mathrm{E_M^p}\right\} = \mathrm{onehot}(1,2, \cdots, \mathrm{M}) \times \mathrm{M_{wpe}} \quad (2\text{-}2)$$

Where $\mathrm{E_i^p}$ is the position of the corresponding word vector onehot. The function transforms the numbers into one-hot vectors of the maximum input length of the model. The embedding matrix $\mathrm{M_{wpe}}$ parameters are also optimized during the training process to obtain the best positional encoding.

2.3.4 Vision Encoder Module

In order to integrate information at the embedding level, fMRI signals from different ROIs within the visual cortex should be converted into a single embedding. The present model uses a gated recurrent unit (GRU) to achieve this goal.

Gated recurrent units are a variant of recurrent neural networks (RNNs) (Zaremba, Sutskever, & Vinyals, 2014). Similar in principle to the Long Short-Term Memory (LSTM) model (Graves & Graves, 2012), which also aims to solve the gradient problem in long-term memory and backpropagation (Cho et al., 2014).

After the decoding task acquires fMRI signals from different primary visual areas, the model splices the signals from each ROI and also the ROI index into the embedding dimension to form a sequence of ROI features in the embedding space. Assuming that the decoding task involves the calculation of N visual regions, then the visual region feature sequence $\{x^1, x^2, x^3, \ldots, x^N\}$ can be obtained as an input sequence to the GRU model, the GRU model computes the corresponding hidden activation $\{h^1, h^2, h^3, \ldots, h^N\}$ and outputs a sequence of vectors $\{y^1, y^2, y^3, \ldots, y^N\}$ as the output sequence to the GRU model.

The formula for the visual coder is given in Equation (2-3):

$$\{\mathrm{y}^1, \mathrm{y}^2, \ldots, \mathrm{y}^\mathrm{N}\} = GRU(\{\mathrm{x}^1, \mathrm{x}^2, \ldots, \mathrm{x}^\mathrm{N}\}; \theta^{GRU}) \qquad (2\text{-}3)$$
$$\mathrm{E}^\mathrm{v} = \mathrm{y}^\mathrm{N}$$

where x^n is the n-th- feature vector in the input brain signal feature sequence, and y^n is the hidden state vector and output vector of the n-th step loop, and θ^{GRU} is the parameter set of the gated loop unit. The output of the last loop step y^N will contain the information of the antecedent feature sequence, we select y^N as the brain signal embedding feature E^b as the brain signal embedding features.

The formula for the GRU model is presented in equation (2-4):

$$\begin{aligned} r^n &= \sigma(W_r x^n + U_r h^{n-1} + b_r) \\ z^n &= \sigma(W_z x^n + U_z h^{n-1} + b_z) \\ \tilde{h}^n &= \tanh(W_h x^n + U_h(r^n \odot h^{n-1}) + b_h) \\ h^n &= z^n \odot h^{n-1} + (1 - z^n) \odot \tilde{h}^n \\ y^n &= \sigma(W_o h^n + b_o) \end{aligned} \qquad (2\text{-}4)$$

where x^n is the first vector of the input ROI feature sequence n vector, h^n is the output vector of the n-th step, and r^n is the state of the reset gate, z^n is the state of the update gate. W, U and b are the weight matrix and bias vector parameters to be learnt, and σ represents the element-by-element sigmoid activation function.

In this way, each vector in the output sequence has access to information from all contexts (prior ROIs in the sequence).

2.3.5 Network Encoder Module

In order to fuse the global information of the brain without introducing too much noise and irrelevant signals, this paper chooses not to use the whole brain signals directly. Instead, fMRI signals from the whole brain signals are used as the context and these signals are utilized to compute cross-attention with the visual area signals. This method hopes to further enhance the encoded semantic information with the global information of the brain as the reference.

The network encoder uses a multi-head cross-attention mechanism to facilitate feature selection. First, in this paper, fMRI signals from all regions of the brain are sampled into the embedding dimension using linear interpolation. It is assumed that the brain has a total of M cortical regions, of which N are visual cortical regions, then the visual region feature sequences $\{x^1, x^2, x^3, \dots, x^N\}$ and the whole brain region feature sequence $\{o^1, o^2, o^3, \dots, o^M\}$ can be obtained as input sequences to the network encoder. The two vector sequences are input in the form of a matrix to the H attention paths. The value of H is the number of heads of the multi-head cross-attention model, which is specified manually. In this model, we set the number of heads to 8. Subsequently, the features of the visual area are input to the linear layers k and v and the features of other visual regions are input into the linear layer q. The output of the linear layer is multiplied with the transposed output of the linear layer k, which is then multiplied by the transposed output of the linear layer and passed through the Scale layer and the Softmax layer. Finally, the output of the Softmax layer is multiplied by the output of the linear layer v.

The formula for the calculation of the i-th cross attention path is shown in Equation (2-5):

$$\begin{aligned} Q^i &= q^i(O) \\ K^i &= k^i(X) \\ V^i &= v^i(X) \\ A^i &= Softmax\left(Scale\left(Q^i \times {K^i}^T\right)\right) \times V^i \end{aligned} \tag{2-5}$$

Where X is the matrix created by splicing the feature vectors of the visual regions, and O is the matrix created by splicing the feature vectors of all regions of the whole brain. The outputs of the different attentional paths are then concatenated and fed into the linear layer for dimensionality reduction. After averaging the features of the different brain regions, the model can obtain the final network embedding. The formula for this step is shown in Equation (2-6).

$$E^{net} = Avg\left(Concat(A^1, A^2, A^3, \cdots, A^H)\right) \quad (2\text{-}6)$$

2.3.6 GPT Decoder Module

The GPT-2 model is a generative language model developed by OpenAI. It has the ability to generate a large number of text sequences and adapt to different text input styles and contents. In addition, GPT-2 shows versatility in performing a variety of natural language processing (NLP) tasks, including classification, information extraction, and language generation. Architecturally it consists of a multi-layer self-attention mechanism and a feedforward neural network.

Each layer in the decoder performs two main operations: self-attention and feedforward neural networks. The self-attention mechanism allows the model to focus on different locations in the input sequence, capturing dependencies and correlations between markers. It assigns different weights to different tokens based on the correlations between them, allowing the model to focus on the most informative parts of the input.

In the self-attention mechanism, the decoder in GPT-2 uses multi-head attention. This means that it performs the attention computation in parallel using multiple sets of learned attention weights, thus enabling the model to capture different types of relations and dependencies. After the self-attention step, the decoder applies a feedforward neural network to each labeled representation. The network consists of two linear transformations with a nonlinear activation function introduced in the middle, such as the Gaussian Error Linear Unit (GELU). It helps to capture complex patterns and relationships in the encoded representations.

The GPT Decoder receives features from the Word Embedding Module, Position Embedding Module, and brain signal Encoders and uses them as context for decoding text generation for the appropriate subjects and the appropriate tasks. The caption decoder adopts the structure of the GPT generative language model, i.e., the decoder part of the Transformer. The embedded features from each upstream module are firstly feature fused to integrate the textual control information from Prompt with the brain signal information, as shown in Equation (2-7):

$$\{\mathrm{E}_1, \mathrm{E}_2, \cdots, \mathrm{E}_\mathrm{M}\} = \left\{\mathrm{E}_1^\mathrm{w} + \mathrm{E}_1^\mathrm{p} + \mathrm{E}^\mathrm{b} + \mathrm{E}^\mathrm{net}, \mathrm{E}_2^\mathrm{w} + \mathrm{E}_2^\mathrm{p} + \mathrm{E}^\mathrm{b} + \mathrm{E}^\mathrm{net}, \cdots, \mathrm{E}_\mathrm{M}^\mathrm{w} + \mathrm{E}_\mathrm{M}^\mathrm{p} + \mathrm{E}^\mathrm{b} + \mathrm{E}^\mathrm{net}\right\} \quad (2\text{-}7)$$

Subsequently, the embedded feature sequences are fed into a multilayer Transformer module structured as a multi-head self-attention, and finally the predicted probability of the target word is obtained through a feedforward neural network, as shown in Eq. (2-8):

$$\mathrm{h}_0 = \{\mathrm{E}_1, \mathrm{E}_2, \cdots, \mathrm{E}_\mathrm{M}\}$$

$$h_l = \text{transformer\_block}(h_{l-1}) \quad (2\text{-}8)$$

$$P(u) = \text{softmax}(h_L W_e^T)$$

Where h_l is the sequence of features output by thel feature sequence output by the layer Transformer module, and L is the number of layers of the Transformer module, and W_e is the mapping matrix of feature vectors to glossary word indexes, and $P(u)$ is the output word u probability of the output word.

Given the contextual features, the GPT-based caption decoder optimizes the model by maximizing the conditional probability of generating the correct word and achieves continuous sentence text generation through an autoregressive inference process. The formulas are shown in Eq. (2-9):

$$L = \sum_i \log P(U_i | E_i, \dots, E_M, U_1, \dots, U_{i-1}; \theta) \quad (2\text{-}9)$$

Where L is the objective function, i.e., the conditional probability to be maximized, and U_i is the i-th feature vector of the image describing text, and θ is the parameter set of the caption decoder. The GPT-2 model repeats the process of self-attention and feedforward neural networks at each layer of the decoder. Each layer builds on the representation generated in the previous layer, allowing the model to continuously refine its understanding of the input and generate more accurate and contextually relevant text. Finally, the framework of this model will generate two types of outputs based on the cues and brain signals, a category text in the form of words and a semantic description text in the form of sentences.

2.4 Decoding experiment results

In this paper, a series of experiments were conducted to validate the ability of the proposed model to decode high-level semantic information in the brain. In order to localize and segment the different cortical regions of the brain, I used the HCP Multimodal Partitioning 1.0 Atlas (HCP-MMP1) as an atlas for selecting brain regions. The HCP-MMP1 atlas uses multimodal magnetic resonance images from the Human Connectivity Project (HCP) and objective, semi-automated neuroanatomical methods to depict 180 regions in each hemisphere. These regions are based on precisely aligned population averages of 210 healthy young adults and are defined by significant changes in function, connectivity, or topography (Glasser et al., 2016). In my experiments, I merged 360 regions from the left and right brain and used 180 brain regions containing parts of the left and right brain.

Direct prediction using raw features from all cortical regions can lead to a number of problems, including computational difficulties caused by hardware memory usage and a large amount of extraneous noise that is difficult to eliminate. Therefore, the experiment input features from all 180 brain regions covering the entire cortex into a network encoder block. In the network encoder block, features from all regions in the brain will be used as contextual guides for fusion with visual area signals for augmentation of the global signal, which helps the model to obtain an approximate representation of the global information of the brain.

In addition, in the visual encoder block, the experiments used original features of visually relevant areas, including a selection of visual areas that included the ventral and dorsal visual pathways, the MT+ complex and its neighbors, as well as a number of areas associated with visual information from higher sensory cortex. Details of the cortical regions used in this paper are shown in Table 2-1.

Table 2-1 Visual Area Information Table

Brain Regions	Full Name	Lobe of the brain	Cortex	Subject 01 Number of voxels	Subject 02 Number of voxels	Subject 05 Number of voxels	Subject 07 Number of voxels
V1	Primary Visual Cortex	Occipital Lobe	Primary Visual	4308	3166	2996	3284
V2	Second Visual Area	Occipital Lobe	Early Visual	2916	2399	2355	2147
V3	Third Visual Area	Occipital Lobe	Early Visual	2096	1705	1720	1560
V3A	Area V3A	Occipital Lobe	Dorsal Stream Visual	652	560	547	488
V3B	Area V3B	Occipital Lobe	Dorsal Stream Visual	189	205	148	139
V3CD	Area V3CD	Occipital Lobe	MT+ Complex and Neighboring Visual Areas	259	272	199	185
V4	Fourth Visual Area	Occipital Lobe	Early Visual	1311	1053	1181	964
V4t	Area V4t	Occipital Lobe	MT+ Complex and Neighboring Visual Areas	217	167	206	212
V6	Sixth Visual Area	Occipital Lobe	Dorsal Stream Visual	394	376	346	377
V6A	Area V6A	Occipital Lobe	Dorsal Stream Visual	223	223	232	213
V7	Seventh Visual Area	Occipital Lobe	Dorsal Stream Visual	186	246	181	194

V8	Eighth Visual Area	Occipital Lobe	Ventral Stream Visual	330	241	298	260
VMV1	VentroMedial Visual Area 1	Occipital Lobe	Ventral Stream Visual	324	233	281	203
VMV2	VentroMedial Visual Area 2	Occipital Lobe	Ventral Stream Visual	245	161	264	158
VMV3	VentroMedial Visual Area 3	Occipital Lobe	Ventral Stream Visual	203	285	231	251
VVC	Ventral Visual Complex	Temporal Lobe	Ventral Stream Visual	482	592	472	433
DVT	Dorsal Transitional Visual Area	Occipital Lobe	Posterior Cingulate	395	501	424	430
FFC	Fusiform Face Complex	Temporal Lobe	Ventral Stream Visual	747	859	540	557
FST	Area FST	Occipital Lobe	MT+ Complex and Neighboring Visual Areas	465	393	266	284
IPS1	IntraParietal Sulcus Area 1	Parietal lobe	Dorsal Stream Visual	365	489	304	274
LO1	Area Lateral Occipital 1	Occipital Lobe	MT+ Complex and Neighboring Visual Areas	151	168	131	158
LO2	Area Lateral Occipital 2	Occipital Lobe	MT+ Complex and Neighboring Visual Areas	304	220	234	229
LO3	Area Lateral Occipital 3	Occipital Lobe	MT+ Complex and Neighboring Visual Areas	218	250	236	234
MST	Medial Superior Temporal Area	Occipital Lobe	MT+ Complex and Neighboring Visual Areas	224	221	216	252
MT	Middle Temporal Area	Occipital Lobe	MT+ Complex and Neighboring Visual Areas	212	246	151	229
PCV	PreCuneus Visual Area	Parietal lobe	Posterior Cingulate	474	387	493	310
PH	Area PH	Temporal Lobe	MT+ Complex and Neighboring Visual Areas	716	849	582	607

PIT	Posterior InferoTemporal complex	Occipital Lobe	Ventral Stream Visual	328	259	334	256
STV	Superior Temporal Visual Area	Parietal lobe	Temporo-Parieto-Occipital Junction	482	516	432	475

2.4.1 Category Decoding Experiment

Depending on the different cues designed for different tasks, the framework proposed in this paper can perform text description decoding and category decoding tasks on a single model.

For the category decoding task, the model predicts the category labels of the visual stimuli received by the subject based on the fMRI signal and generates them in text form. The visual stimuli were taken from the COCO dataset, which contains visual stimuli labeled as person, vehicles, outdoors, animals, accessories, sports, kitchen, food, furniture, electronics, household appliances, and indoors in 12 main categories.

After each stage of model training throughout, the experiment tested the model's category-aware accuracy on a test set. The model reached a stable performance at the third epoch during the training of the model. The category-aware accuracy of the model at different training stages is shown in Figures 2-3.

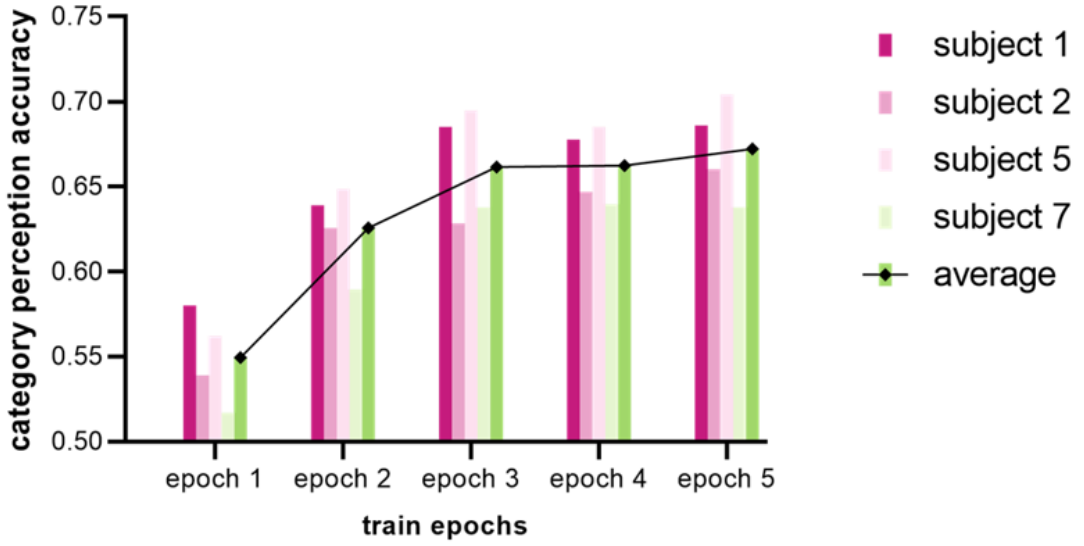

Figure 2-3 Classification accuracy of the test set at different stages of training

The classification accuracy of the fifth epoch alone is 67.21%, which is significantly higher than the random level of 12 classifications. However, due to the nature of GPT as an autoregressive text generator, the predictions of the model are not always one of the 12 labels

but may be any permutation or combination of words in the entire vocabulary. As a result, the level of stochasticity in the classification task based on GPT text generation is actually much lower than that of the 12 classification task. Nevertheless, the model still achieved good performance.

The confusion matrix of the category-aware task performed by the proposed model on four participants' data is shown in Figure 2-4. I categorize outputs other than standard labeled text as 'others'. From the confusion matrix, we can observe that categories with semantically informative containment relationships have a higher probability of being predicted, even in the case of prediction errors. For example, for a visual stimulus triggered by an image of "kitchen", the output labels of the model will focus on "furniture" and "food". This suggests that the proposed model captures and recognizes similar patterns of brain activity triggered by similar visual stimuli and associates them with semantically similar language.

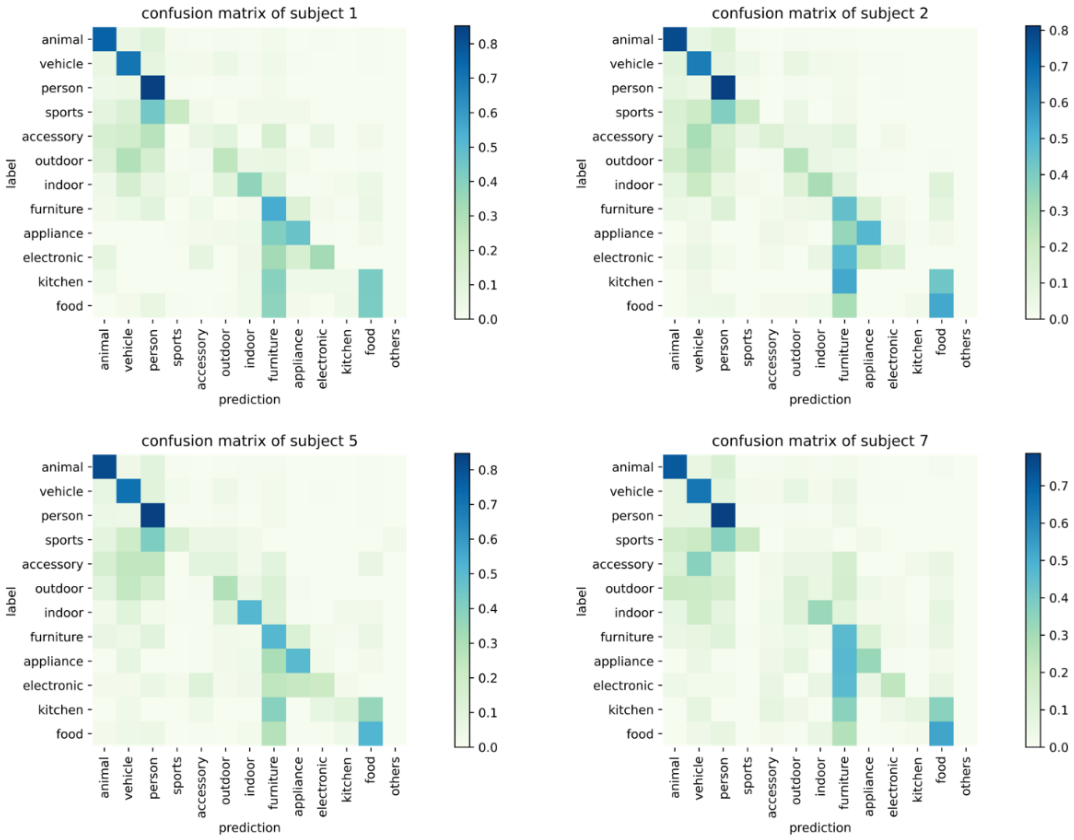

Figure 2-4 Confusion matrix for category-aware experiments

2.4.2 Caption Decoding Experiment

For the caption decoding task, the proposed model accepts the "caption decoding" cue and the fMRI signals to be decoded and then generates text to describe the visual stimulus. Part of

the results of the proposed framework for language decoding using fMRI signals from the NSD dataset are shown in Figures 2-5.

Manually annotated caption:

1. A man on a surf board riding a wave in the ocean.
2. A person riding a wave on top of a surfboard.
3. A guy surfing on a wave in the ocean.
4. A man in a wet suit surfing on a wave.
5. A person on a surf-board riding an ocean wave.

Predicted caption:

Subject 01: A man in a wet suit riding a wave on a surfboard.
Subject 02: A person sitting on a surfboard in the water.
Subject 05: A man riding a wave on top of a surfboard.
Subject 07: A man riding a surfboard on top of a wave.

Manually annotated caption:

1. A zebra standing in tall brown grass in front of trees.
2. A zebra stands in a field of dried shrubs.
3. A single zebra standing in a tan veldt.
4. A zebra standing in a dry field of grass.
5. A lone zebra stands in the dry savannah.

Predicted caption:

Subject 01: A zebra standing in a grassy field next to a pile of rocks.
Subject 02: A zebra grazing on grass in a field.
Subject 05: A zebra standing in a grassy field with trees in the background.
Subject 07: A zebra standing in a grassy field next to trees.

Manually annotated caption:

1. A very tall clock tower with two clocks on it.
2. A tower with a clock on it with a sky background.
3. A view of a clock tower from the street.
4. A very large building with a clock on top.
5. A big building with a clock built inside of it.

Predicted caption:

Subject 01: A large building with a clock on the front.
Subject 02: A large clock tower with a statue on top.
Subject 05: A large clock tower with a clock on each of it's sides.
Subject 07: A large clock tower with a clock on each of it's sides.

Manually annotated caption:

1. A white plate topped with spaghetti and broccoli.
2. A meal of noodles and broccoli on a plate.
3. A dish of broccoli and spaghetti with red sauce.
4. Pasta with broccoli and tomatoes on a white plate.
5. A plate of food has noodles and broccoli.

Predicted caption:

Subject 01: A plate of broccoli, cauliflower, and carrots with ranch dip.
Subject 02: A plate of food with broccoli, potatoes and meat.
Subject 05: A plate of broccoli and shrimp stir fry.
Subject 07: A plate of broccoli and noodles with a fork.

Manually annotated caption:

1. A woman holding a racquet on a tennis court.
2. A girl on a tennis court hitting a tennis ball with racket.
3. A girl in a pink t-shirt and white tennis shoes holds out her racket.
4. Woman wearing pink and black swinging at a tennis ball.
5. A female tennis player holding a tennis racket about to hit the tennis ball.

Predicted caption:

Subject 01: A young boy holding a tennis racquet on a tennis court.
Subject 02: A young man holding a tennis racquet on top of a tennis court.
Subject 05: A woman holding a tennis racquet on a tennis court.
Subject 07: A man holding a tennis racquet on a tennis court.

Figure 2-5 Partial results of the caption decoding task for four subjects (a)

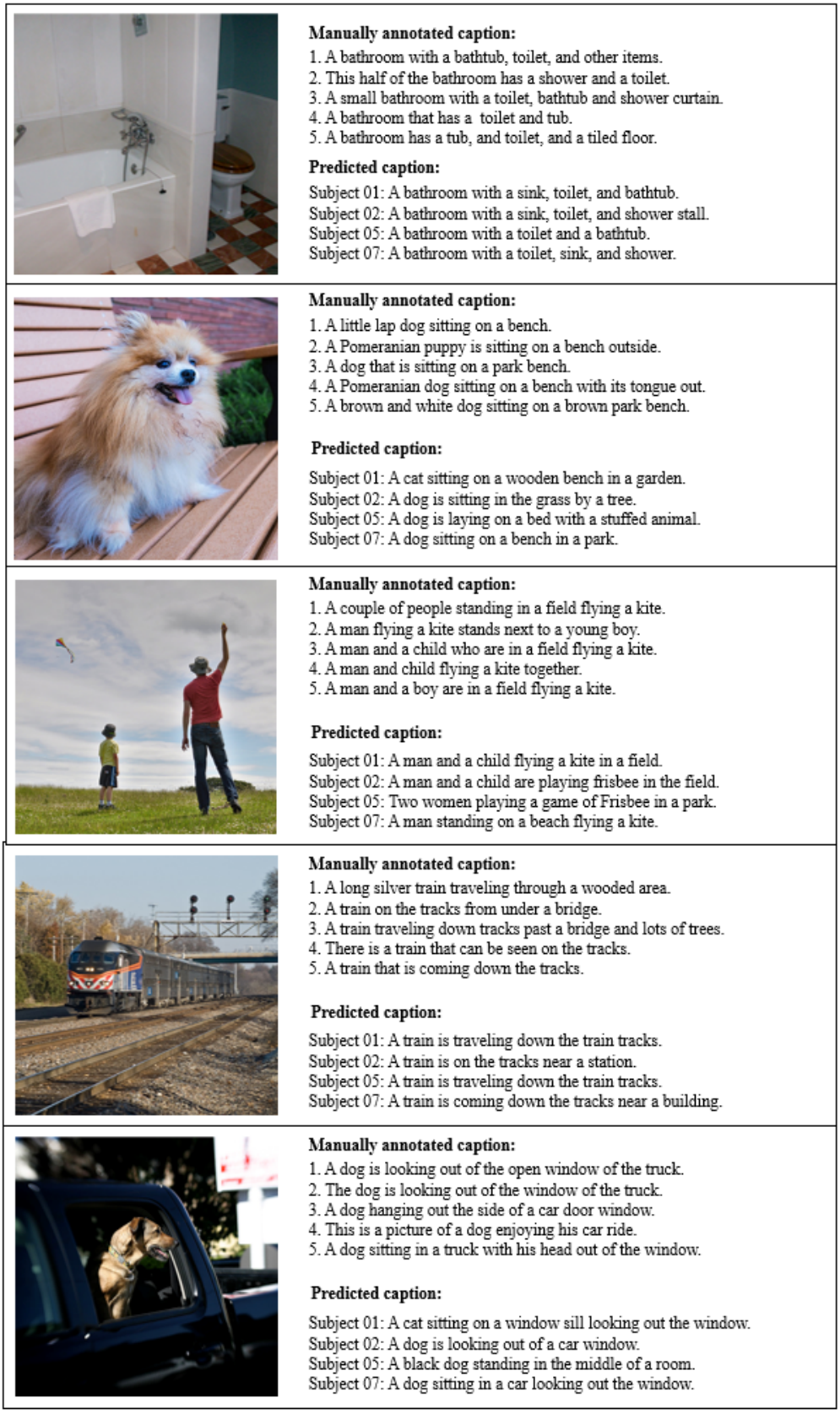

Figure 2-5 Partial results of the caption decoding task for four subjects (b)

The fMRI text descriptions generated by the model are semantically similar to the visual stimulus content and manually annotated text descriptions. The generated text even describes the information that has not yet been described in the manually annotated text descriptions, such as color information, although there are still some biases in understanding the semantic relationships.

In order to quantitatively assess the performance of the proposed method in text description decoding, I choose Bilingual Evaluation Understudy (BLEU) (Papineni, Roukos, Ward, & Zhu, 2002), Recall-Oriented Under-study for Gisting Evaluation (ROUGE),(C.-Y. Lin, 2004), and Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Banerjee & Lavie, 2005) as evaluation metrics.

These metrics will measure the semantic similarity between the fMRI text description generated by the model and the manually annotated text, so that the higher the value of the metric, the closer the description text generated by the model is to the human description.

The distribution of semantic similarity metrics between model-generated text descriptions and manually annotated text descriptions with respect to random sampling results is shown in Figures 2-6. Compared with random sampling, the semantic similarity between the model-generated text descriptions and manually annotated text descriptions is significantly higher, which fully demonstrates the validity of the method proposed in this paper in the task of fMRI text description generation.

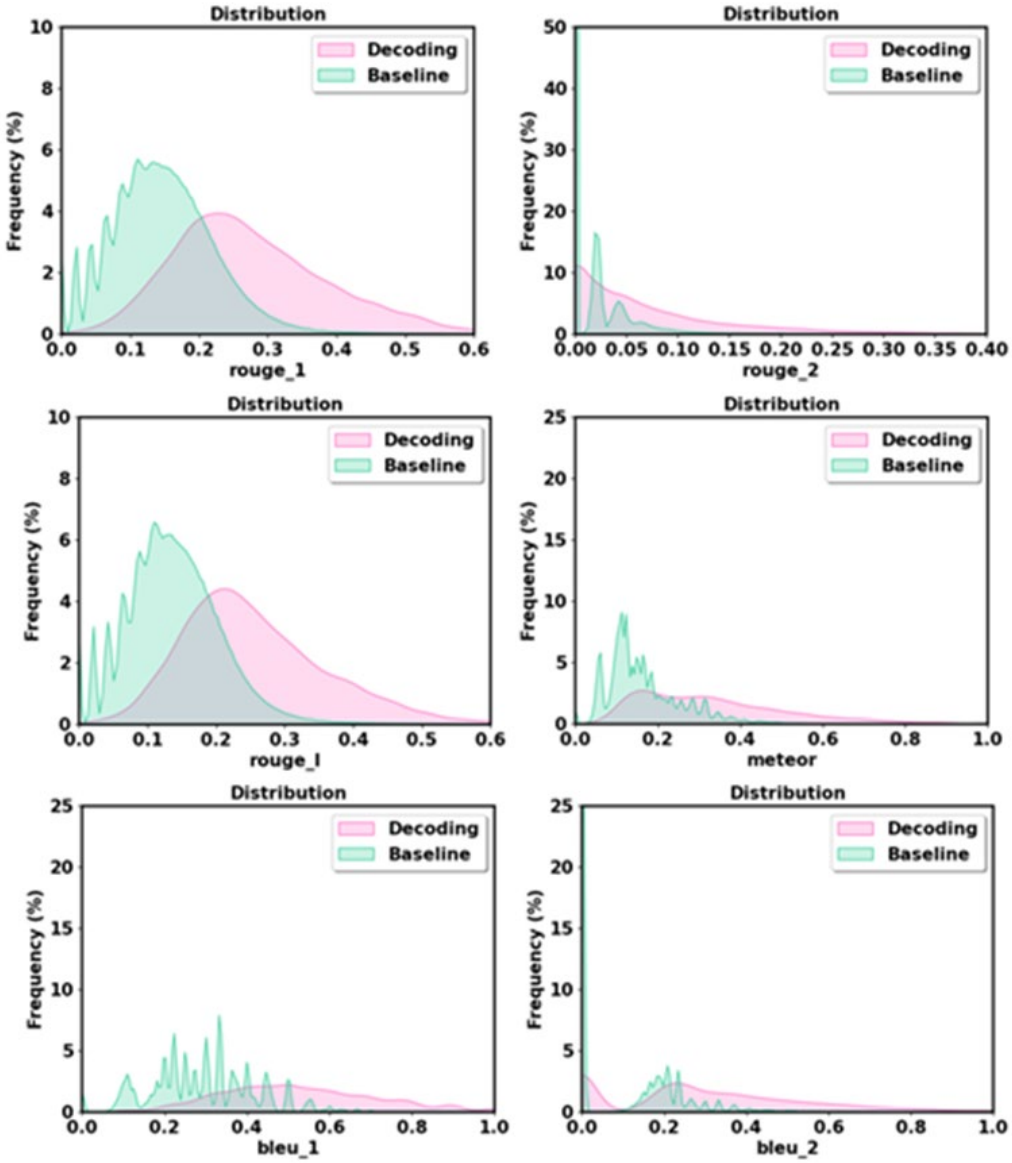

Figure 2-6 Comparison of the distribution of descriptive text and random level fidelity metrics predicted by the model

In order to better measure the performance of the model proposed in this paper, I compare it with the results of three recent models that perform caption decoding tasks, namely Brain Captioning (Ferrante, Ozcelik, et al., 2023), DreamCatcher (Chatterjee & Samanta, 2023), and UniBrain (Mai & Zhang, 2023), which also performed the fMRI caption decoding task on the NSD dataset. The comparison results are shown in Table 2-2.

Comparative results show that the proposed model outperforms other current caption decoding models in all metrics, which proves the excellent performance of our model in the fMRI text description decoding task.

Table 2-2 Comparison results of DecodeGPT with several other state-of-the-art caption decoding models on the caption decoding task (bolded numbers represent the highest performance achieved in each metric compared to the other compared models)

Indicators	Ours	Brain Captioning	Dream Catcher	UniBrain
Meteor (Subject 01)	**0.353**	0.305	0.323	
Meteor (Subject 02)	**0.333**	0.298	0.308	
Sentence (Subject 01)	**0.470**	0.447	0.451	
Sentence (Subject 02)	**0.437**	0.418	0.422	
Meteor	**0.342**			0.169
Rouge-1	**0.283**			0.245
Rouge-L	**0.262**			0.222

2.4.3 Ablation Experiment

To verify the effectiveness of encoding information from the whole brain, I also performed ablation experiments on network encoders. The term "visual" in the table refers to the model that retains only the visual encoder and the cue embedding block. The results for caption decoding are shown in Table 2-3, while the results for classification decoding are shown in Table 2-4.

Table 2-3 Comparison of results in the category decoding task for models that include the network encoder block and those that do not include it (bold numbers indicate the best performance achieved in each metric in the models compared)

Subject	Visual+Other	Visual+Noise	VisualOnly	Baseline
Subject 1	**0.678**	0.637	0.667	0.083
Subject 2	**0.647**	0.599	0.635	0.083
Subject 5	**0.685**	0.658	0.684	0.083
Subject 7	**0.640**	0.610	0.607	0.083
Average	**0.662**	0.626	0.648	0.083

Table 2-4 Comparison of results in the caption decoding task between models that include the network encoder block and those that do not include it (bold numbers indicate the best performance achieved in each metric in the models compared)

Metrics	Visual+Other	Visual+Noise	VisualOnly	Baseline
Rouge-1	**0.283**	0.278	0.278	0.148
Rouge-2	**0.069**	0.065	0.065	0.013
Rouge-L	**0.262**	0.258	0.257	0.138
Meteor	**0.342**	0.328	0.335	0.167
BLEU-1	**0.535**	0.531	0.529	0.294
BLEU-2	**0.295**	0.286	0.288	0.072

In addition to the "VisualOnly" representation of a model containing only visual pathways, we also compared the decoding metrics with random levels. In the category decoding task, the "Baseline" is derived from the chance level of the 12 categories. The "Baseline" for caption decoding performance assessment was calculated by randomly sampling each sentence in the test dataset, randomly selecting another 1024 sentences from the dataset, and calculating the average sentence similarity metrics between the selected sentences and these 1024 samples. Finally, the experiment averages the results of all sentences in the test data set at the "random sampling" level.

As shown in Tables 2-4, for the text description decoding task, the model with the added network encoder outperforms the model with only visual information on all metrics and is much higher than the baseline computed by random sampling.

For the classification decoding task, Tables 2-3 show that the addition of the network encoder improves classification accuracy for the majority of subjects, as well as the average classification accuracy at all subject levels. This demonstrates, again, the contribution of encoding whole brain information to the decoding task.

2.5 Summary and Discussion

In this study, we propose a new universal visual decoding model, DecodeGPT, which is capable of multi-subject and cross-subject decoding within a single model and, at the same time, still has a decoding performance comparable to other state-of-the-art decoding models in each task for each subject. In order to achieve multi-subject and cross-subject decoding, we design Prompt text instructions containing "subject + task" information. We fine-tune GPT-2 based on the Prompt text instructions and the corresponding fMRI sample dataset so that it is

able to process potential features after adding brain signal features and decode the corresponding subject and the corresponding task. The output text of the GPT is labeled in the form of words if Prompt instructions for the categorization task are received. The output text of the GPT is natural language descriptions of the content of the visual stimuli in the form of sentences if Prompt instructions for the text description task are received. When the model performs the category decoding task, we directly generate predicted labeled words through GPT, which are 12 semantic labels including Person, Vehicles, Outdoor, Animals, Accessories, Sports, Kitchen, Food, Furniture, Electronics, Household Appliances, and Indoor. In the classification task, the traditional neural network-based classification model decodes the result from the maximum value of the output layer after the Softmax function. Hence the chance level is 8.33%. However, the decoding result of the classification model based on the GPT language model comes from the generated text of the language model, and since the generated content of the GPT is of any length with the text containing any words, the chance level of using the language model for the classification task is close to 0 probability. Despite this, the proposed model still shows a good classification performance for the classification task for each of the subjects, which is far beyond the chance level of classification. Moreover, in the confusion matrix, the proposed method shows a greater probability of confusion for semantically similar categories, such as between "kitchen" and "food", "appliances" and "electronics". This suggests that the model proposed in this paper successfully establishes a meaningful and generalized mapping between each subject's brain visual stimulus-response patterns and semantic concepts. This phenomenon may, to some extent, reflect the way the human brain works when processing semantic information, i.e., there is a degree of ambiguity and confusion in similar semantic domains. Therefore, the performance of the model in these domains is consistent with the characteristics of human cognition, which further validates the accuracy and generalization ability of the model to understand semantics. In the caption decoding task, the model proposed in this paper generates natural language text descriptions that can accurately reflect the content of the original visual stimulus images, and there is a high level of semantic similarity between the generated text descriptions and descriptions derived from human language, which is far beyond the random level in terms of semantic similarity. Compared with some current works that perform caption decoding, the method proposed in this paper has a better performance on the caption decoding task, and this performance improvement may be brought about by the GPT language model obtained based on a large amount of corpus pre-training. In terms of the

linguistic similarity evaluation metrics of the generated text, the proposed model slightly outperforms the DreamCatcher method based on the pre-trained GPT word embedding space and outperforms other caption decoding models that do not use pre-trained GPT information. Finally, the ablation experiments for the network encoder module in this paper show that the network encoder designed and introduced in this paper can effectively improve the performance of the model on the classification decoding task as well as the caption decoding task.

However, confusion between certain labels still exists in classification tasks. This may be attributed to the imbalance in the number of samples of different labels in the dataset and the limited distinguishability of the labels themselves based on image content. In addition, the dimensionality of the representation of the brain signals is limited by the relatively low dimensionality due to the size constraints of the models used. This limitation may affect our ability to decode more detailed and accurate semantic information. Nevertheless, the model proposed in this paper has excellent scalability. By employing different prompt cue word designs and substituting different encoding paths, the model can theoretically be extended to more kinds of text generation-based decoding tasks. Thus, the proposed architecture also contributes to the realization of more flexible and versatile brain-computer interfaces.

3. A cross-subject universal decoding method for data migration

3.1 Introduction

The field of Brain-Computer Interface (BCI) continues to evolve with advances in pattern recognition and neural signal acquisition techniques. In 1924, Hans Berger recorded neural activity for the first time by means of Electroencephalogram (EEG), opening the door to exploring the mechanisms of neurological diseases. Today, a number of non-invasive and invasive neural signal acquisition techniques have been developed, including functional magnetic resonance imaging (fMRI) (Logothetis, 2008), functional near-infrared spectroscopy (fNIR) and cortical electroencephalography(EEG)(Ferrari & Quaresima, 2012) and electrocorticography (ECoG) (Buzsáki, Anastassiou, & Koch, 2012). These tools enable brain decoding tasks such as identifying and classifying unique brain responses to different types of stimuli and activities. The high spatial resolution and accessibility of functional MRI make it ideally suited for studying patterns of neural activity in specific regions of the brain.

In fact, aside from the advancements in brain signal acquisition methods, decoding models used for pattern recognition of brain signals also face challenges and are continuously evolving. The development of support vector machines (Platt, 1998) and the emergence of various machine learning models based on deep neural networks have led to continuous improvements in the performance of brain-machine interface models (Khademi, Ebrahimi, & Kordy, 2023).

However, models in the field of deep learning have always faced the problem of data scarcity, and brain-computer interfaces based on deep learning models for pattern recognition are no exception. One of the main challenges in the field of brain decoding is the scarcity of data due to the high cost of collecting brain signals. Willett et al. have implemented a remarkable brain-computer interface application that converts the imagined process of "handwriting" in the brain into text on a screen, thus helping a patient who is incapacitated due to paralysis below the neck to input information by simply imagining writing letters in his head (Willett, Avansino, Hochberg, Henderson, & Shenoy). Although the experiment collected more than 30,000 neural signals from the patient's imagined writing of letters over a total acquisition time of nearly eight hours, the amount of data collected was still very limited for model training, and the decoding model still faced a serious challenge of overfitting.

Another challenge is that models trained using data from one subject cannot be directly applied to data from other subjects, which leads to significant performance degradation (Wen et al., 2018). Therefore, in current neural decoding work, training decoding models for each

subject individually is still the dominant choice. For example, in the multi-task decoding model proposed by Mai et al. (Mai & Zhang, 2023) , they need to train subject-specific decoding models for each of the four subjects in the NSD dataset. This means that if we want to apply the decoding model to new subjects in a brain-computer interface, we need to collect a large amount of new data and train a large number of new decoding models that are subject-specific. As a result, the cost of acquiring decoding models will become non-negligible as the number of subjects continues to increase.

Establishing a method to achieve cross-subject data migration can help solve both problems. By migrating all subjects' data into a common feature space, we can build a common decoding model shared by all subjects. Migrating new subjects' data to the common feature space using the subject data migration technique will make the new subjects' data available for existing decoding models, thus avoiding the need to train new decoding models that are subject-specific. Moreover, on the univeral decoding model, new subjects can make use of previous subjects' data, which is equivalent to enlarging the size of the new subjects' dataset and alleviating the overfitting problem to some extent.

Although feature alignment can align fMRI data from different subjects to a common feature space, cross-subject data migration cannot be achieved by simply using these feature alignment methods. This is because the common space obtained after training is only an optimal solution based on the existing data. This means that if a new subject is added, we have to re-align all previous subjects with the new subject. And since adding a new subject leads to a change in the common space, the neural decoding model also needs to be retrained. Therefore, the data from new subjects cannot be used directly on the previously trained model.

To address this problem, inspired by a new semi-supervised multi-view learning approach (Hu et al., 2021), we propose a new functional alignment method for cross-subject data migration. The innovation of this method compared to traditional feature alignment methods is that it solves a series of problems caused by the increasing number of subjects. First, the method achieves a more stable common feature space compared to traditional feature alignment methods by applying a fixed orthogonal matrix shared among subjects, so that the new subjects' data can be used in the existing decoding model directly without retraining after the data is aligned. At the same time, this method realizes the sharing of data between subjects, which effectively reduces the demand for new subjects' data collection. By adopting the generalized

contrast learning approach, the method also does not rely on temporally aligned subject data, which further reduces the data requirements.

3.2 Experiment data

In the experiments, this paper first uses two publicly available datasets from OpenNeuro. The first dataset is ds000105, which contains spatial sizes of $64 \times 64 \times 40$ of BOLD images acquired from a GE 3T (General Electric, Milwaukee, WI) scanner [repetition time (TR) = 2500 ms, 40 3.5-mm thick sagittal images, field of view (FOV) = 24 cm, echo time (TE) = 30 ms, flip angle = 90]. In ds000105, a total of six subjects viewed a variety of stimuli, with stimulus types including pictures of faces, cats, and five categories of man-made objects (houses, chairs, scissors, shoes, and bottles), as well as a control group of images containing a meaningless random pattern. Each stimulus category has multiple examples, each with 4 images, for a total of 12 different examples. The meaningful stimuli will be repeated, with the repeated stimuli being pictures of the same faces or objects taken from different angles. During the fMRI scans, each subject was scanned during 12 runs. Each run began and ended with a 12-second rest period and consisted of eight stimulus blocks lasting 24 seconds. There was one stimulus block for each category, and they were separated by 12-second rest intervals. Stimuli were presented with a duration of 500 ms and a stimulus interval of 1,500 ms. The ventral temporal lobe (VT) cortex was selected as the region of interest based on previous studies (Hanson, Matsuka, & Haxby, 2004; J. V. Haxby et al., 2001; O'toole, Jiang, Abdi, & Haxby, 2005) .

The second dataset was ds000117, which included 19 subjects (including 8 females and 11 males) aged between 23 and 37 years, all from the MRC Cognition & Brain Sciences Unit participant panel. Each subject was presented with a visual stimulus containing human faces, the facial stimuli consisted of two sets of 300 greyscale photographs each, half of which were famous faces recognizable to most UK adults, and half of which were non-famous faces matched to the famous faces in terms of gender and age, in addition to which the experiment randomly generated confusing faces to serve as control stimuli. The experimental stimuli were first presented with a fixation cross-picture and then followed by either faces of different durations or confusing faces. Each image was presented twice, either immediately or repeated 5-15 stimuli later. To maintain subjects' attention to the face visual stimuli, subjects were asked to indicate whether the images were more symmetrical than average. fMRI data were collected

using a Siemens 3T TIM TRIO scanner and included structural and functional scans of the interleaved slices(Wakeman & Henson, 2015).

3.3 Model

In this paper, we propose a scalable functional alignment model architecture consisting of multiple self-encoders.

The architecture is a set of self-encoders connected by a randomly generated public orthogonal matrix, and each subject is assigned a self-encoder for learning the mapping from the subject's individual space to the public feature space. The intermediate hidden feature of the self-encoder is the target public space, and the encoder input of the self-encoder is the subject individual space feature. The decoder output is the reduced subject individual space feature. The encoder part of the self-encoder models the mapping from the subject individual space to the public feature space for each subject. The intermediate hidden features of each self-encoder are multiplied with a common orthogonal matrix to serve as a constraint fixing the common space across subjects.

Suppose that the experiment has M subjects and that the i-th subject has N_i samples. $X^i = \{x_k^i\}_{k=1}^{N_i}$ is the set of ROI features for the i-th subject, and $Z^i = \{z_k^i\}_{k=1}^{N_i}$ is the set of corresponding one-hot labels. When a dataset of one subject is obtained, this method trains an alignment model for that subject and does not use any other subject's data at all. The architecture of the model is a specially designed autoencoder consisting of two parts: an encoder and a decoder. Both the encoder and decoder use a multilayer perceptron structure, and we can represent the encoder of the i-th subject as f_{en}^i and the decoder of the subject as f_{de}^i . The features of i-th subject in the common feature space can be calculated by Equation (3-1):

$$Y^i = \{y_k^i\}_{k=1}^{N_i} = f_{en}^i(X^i; \theta_{en}^i) \tag{3-1}$$

θ represents the parameters learnt by the neural network during the training process. Then, the decoder reconstructs the features in the individual space based on the latent features (i.e., features on the common space) obtained by the encoder, see equation (3-2):

$$X_{gen}^i = \left\{x_{gen}{}_k^i\right\}_{k=1}^{N_i} = f_{de}^i(Y^i; \theta_{de}^i) \tag{3-2}$$

To ensure that the mapped sample features retain their unique information, this paper uses the encoder loss as a constraint, denoted as$loss_1$, as shown in Equation (3-3):

$$loss_1 = \left\| x_k^i - x_{gen}{}_k^i \right\|_2 \tag{3-3}$$

In addition, this method introduces a fixed, randomly generated orthogonal matrix W, which requires no training and is shared among all subjects. The matrix W transforms the common space features into the labeling space. Due to the orthogonality of the matrix W, it minimizes the intra-class distances of the samples in the common feature space and maximizes the inter-class distances (Fisher, 1936; Sun, Xie, & Yang, 2016). By sharing a fixed matrix across subjects, the present method can fix the aligned common space, thus decoupling the data from different subjects in time and space. In this paper, the predicted features in the labeling space are compared with the corresponding uniquely hot coded labels of the samples and the 2-norm is computed to obtain the comparative learning loss, denoted as $loss_2$, as shown in Eqs. (3-4):

$$\begin{gathered} Z_{pre}^i = \{z\_pre_k^i\}_{k=1}^{N_i} = Y^i \times W = \{y_k^i \times W\}_{k=1}^{N_i} \\ loss_2 = \left\| z_k^i - z_{pre}{}_k^i \right\|_2 \end{gathered} \tag{3-4}$$

The loss function of the alignment model is given by $loss_1$ and $loss_2$ and is balanced by the parameter λ, so the loss function formula is shown in Equation (3-5):

$$loss = (1-\lambda)loss_1 + \lambda loss_2 \tag{3-5}$$

The specific structure of each alignment model in the encoder set is shown in Figure 3-1.

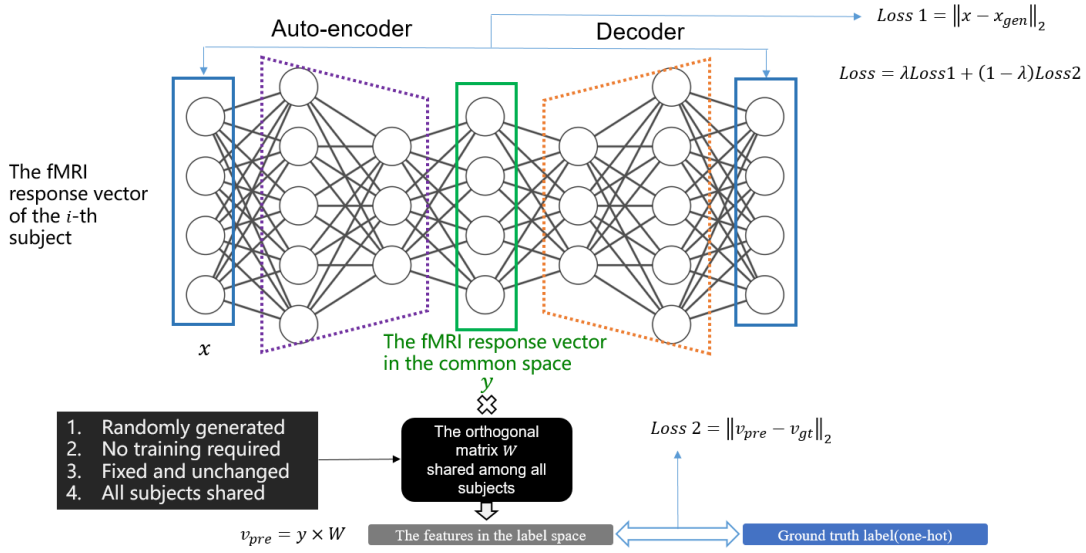

Figure 3-1. Alignment model for the i-th subject

For M individual subjects, this method will train M individual alignment sub-models to transform the features in the individual feature space onto the common feature space. x_k^i is the individual space feature of the k-th sample of i-th subject, and y_k^i is the common space feature of the k-th sample of i-th subject, and $x_{gen}{}_k^i$ is the feature based on y_k^i performing the decoder reconstruction. $z_{pre}{}_k^i$ is the set of features that are obtained through cross product of the y_k^i and orthogonal matrices W. z_k^i is the real labels represented in one-hot form.

Due to the innovative structure of the alignment model proposed in this paper, the method will have the following advantages:

1. The methodology in this paper enables cross-subject data migration, thereby reducing subject-specific data requirements.

2. The approach in this paper decouples the alignment process in time and space, enabling an asynchronous and distributed alignment process.

3. By exploiting generalized constraints on contrast learning rather than pairwise constraints, the alignment model in this paper does not rely on time-aligned fMRI data.

3.4 Experiments in Increasing Subjects Data

3.4.1 Experimental Design

In previous approaches to functional alignment, it is often assumed that in the brains of two individuals receiving the same stimulus (e.g., watching the same full-length film), the cortical response pattern vectors reflect similar information. However, the coordinate systems representing the respective spaces are not aligned. Therefore, in this paper, the space in which the original cortical response pattern vectors of each individual are located is referred to as the individual space. The purpose of feature alignment is to obtain a common feature space shared by all subjects and to map each subject's sample data from the individual space to the common space. This mapping will minimize the differences between the response pattern vectors of different individuals to the same stimulus. This also means that if the individuals involved in the computation are different, the coordinate system of the individual space will be different, and therefore the common space computed based on these individual spaces will be different.

If we wish to implement data migration between subjects using functional alignment, then we need a relatively stable common space. Otherwise, once the data of a new subject is aligned with the data of an existing subject, the original common space will be shifted, resulting in a

decoding model based on the original common space that is no longer applicable to the aligned data after the addition of a new subject.

If the method in this paper can implement a fixed common space, it will be possible to use data from new subjects without having to retrain the downstream decoding model. At the same time, the new subjects' data can also be used to fine-tune other trained models for the subjects' data. The flow of the proposed method in this paper when processing the new subject data is shown in Figure 3-2.

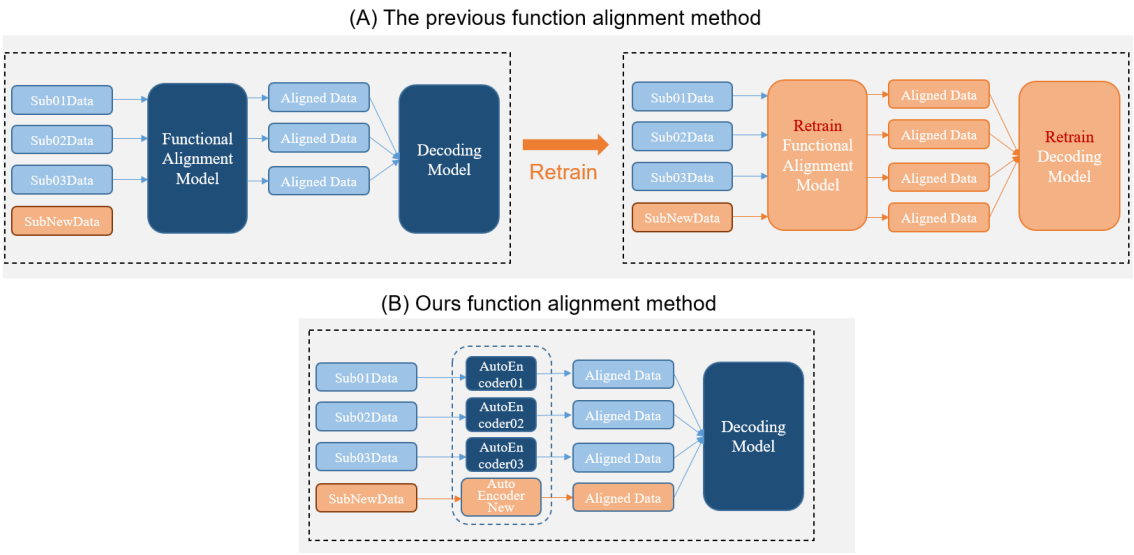

Figure 3-2 Comparison of processes when adding new subject data

(A) When new subject data is added, the proposed method trains a new alignment model for the new subjects, which is used to transform their data from the individual space to a relatively fixed common feature space. The alignment model is trained using only the new subject data, completely independent of other subject data. The fixed public feature space allows direct use of the downstream model for the new subject data and allows us to fine-tune the existing downstream model using the new subject data. (B) When new subject data is added, the traditional HyperAlignment method requires re-aligning the new subject data with all existing data to obtain a new common feature space. Due to the change in the common feature space, the downstream decoding model also needs to be retrained.

In order to verify the extent of the common space shift in different functional alignment methods after adding new subject data, we first conduct experiments to process increasing subject datasets. In the experiments, two publicly available datasets from OpenNeuro are used in this paper: ds000105 and ds000117.

In order to verify the validity of the method in the case of new subject data addition, the following experiments were conducted. First, based on the leave-one-out (LOO) strategy, the data are divided into "new subject data" (including one subject) and "previous subject data" (including other remaining subjects). Then, each dataset is evenly divided into two parts, one is used as a training set to train the alignment model, and the other is used as a test set to subsequently verify the performance of the alignment model. We first train the alignment model using the training set of the "previous subjects' data", and then use the obtained alignment model to convert the test set of the "previous subjects' data" into a common feature space. The converted test set is used to train a decoding model, which we can call "prior decoder". Then, in this paper, we train a new alignment model using a training set that contains all subject data (including "previous subject data" and "new subject data"), and then convert the test set containing all subject data to the public feature space by the new alignment model. In order to measure the effect of the change in the public feature space on the performance of the downstream decoding model, we use the new alignment data containing all subject data as the dataset and use the "previous decoder" as the model to verify the decoding performance of the model.

The experimental flow is shown in Figure 3-3.

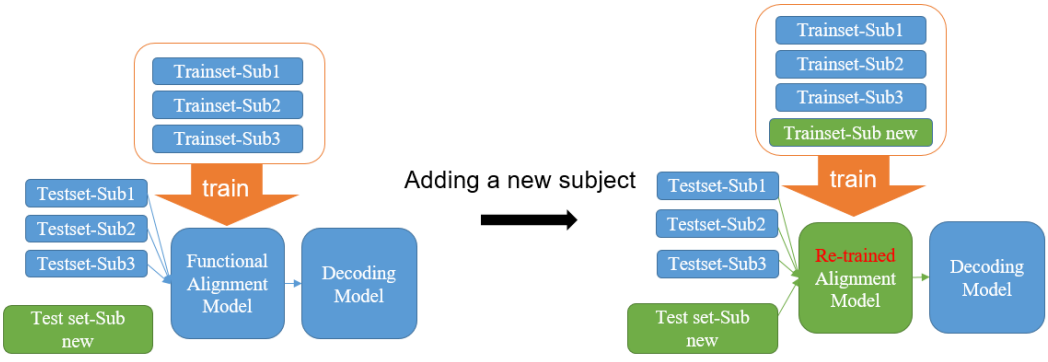

Figure 3-3 Experimental flow of the added subject data for measuring the degree of public space deviation

We use two other functional alignment methods as a control in order to compare the performance of the method we propose with the traditional functional alignment methods. They are Hyperalignment (HA) (James V Haxby et al., 2011) and Regularized Hyperalignment (RHA)(Xu et al., 2012). Hyperalignment does not involve the selection of hyperparameters, while Regularized Hyperalignment includes hyperparameters α. The value of α ranges

between 0 and 1. When α is set to 1, regularized hyper-alignment is mathematically equivalent to hyper-alignment. In subsequent experiments, the hyperparameters of regularized hyper-alignment were derived from the best-performing values obtained from a grid search ranging from 0 to 1.

In order to eliminate the effect of downstream decoding models representing classifiers on the performance of different functional alignment methods, we use exactly the same linear SVM classifiers for different functional alignment methods on each dataset.

3.4.2 Results

Experiments were conducted on three datasets respectively. The correlation results of ds000105 dataset are shown in Fig. 3-4, and the correlation results of ds000117 dataset are shown in Fig. 3-5.

Between-Subject Classification (BSC) accuracy is used to measure the degree of bias in the common space after adding new subjects. Specifically, the Between-Subject Classification Accuracy (BSC) is calculated as follows with the addition of new subjects' data:

a. Leave-one-out: Select one subject as the new subject.

b. Align the data for the remaining N-1 subjects using the alignment method.

c. Train an SVM classifier for the classification task on the aligned N-1 subject data.

d. Introduction of a new subject and alignment of his/her data.

e. Evaluate cross-subject classification accuracy using the re-aligned data from N subjects, including the new subject, as a test set.

f. Repeat steps b-e in a cross-validation manner to obtain an average accuracy.

The performance of the cross-subject classification task is affected due to the use of the decoding model trained on the original public space data on the new public space data. Moreover, the greater the difference between the original public space and the new public space, the lower the accuracy of cross-subject classification will be.

On the ds000105 dataset, we performed 8 classification tasks. The chance level for the cross-subject classification task was 12.50%. The experimental results are shown in Figures 3-4. For the traditional functional alignment method, the cross-subject classification accuracy using the hyper-alignment method is 36.84 ± 9.20%, while the regularized hyper-alignment method corresponds to a classification accuracy of 47.25 ± 14.94%. The method we propose achieved a classification accuracy of 52.50 ± 12.08%.

For the ds000117 dataset, we performed a 3-classification task, resulting in a 33.33% chance level for cross-subject classification. The experimental results are shown in Figures 3-5. Hyper-alignment achieves 68.07 ± 5.54% classification accuracy on this dataset, regularized hyper-alignment achieves 69.02 ± 8.96% classification accuracy, while the method we propose achieves 73.75 ± 9.34% classification accuracy.

For the results obtained using this method and the results obtained using other methods, a two-sample t-test was conducted to compare their means, and despite the small number of subjects, the improvement of our method over the hyper-alignment method was statistically significant (p-value of 0.030 on the ds000105 dataset, and 0.045 on the ds000117 dataset). The results of the statistical test prove that the method we propose is improved relative to both the hyper-aligned and regularized hyper-aligned methods at the average level. Our proposed method performs significantly better than the hyper-align method.

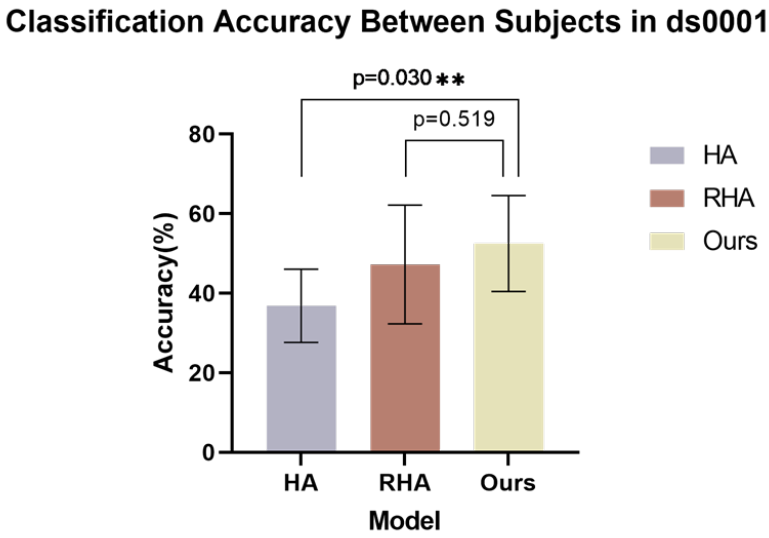

Figure 3-4 Comparison of the performance of our method with Hyper-Alignment (HA) and Regularized Hyper-Alignment (RHA) when processing new subject data on the ds000105 dataset (The error bars represent standard deviation).

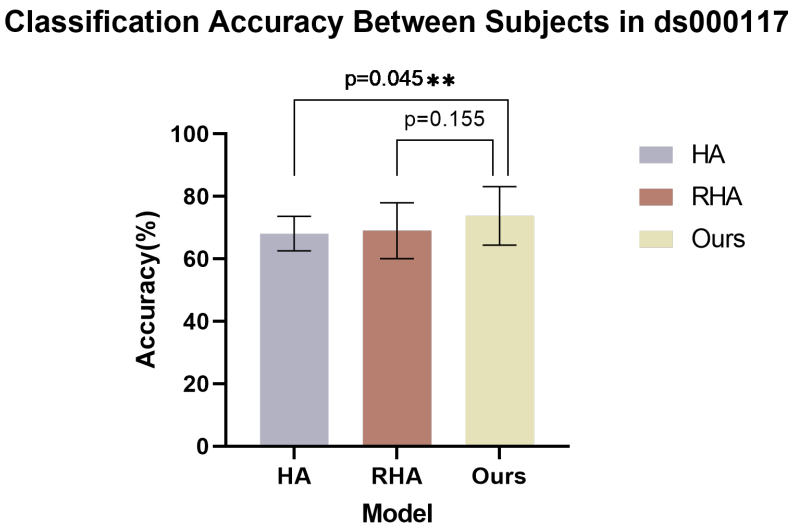

Figure 3-5 Comparison of the performance of our method with Hyper-Alignment (HA) and Regularized Hyper-Alignment (RHA) when processing new subject data on the ds000117 dataset (The error bars represent standard deviation).

Experimental results on the publicly available ds000105 and ds000117 datasets show that the method in this paper achieves better cross-subject classification accuracy than RHA or HA on both 3T BOLD datasets. This implies that the method in this paper effectively suppresses the shift in the common space after adding new subjects relative to the traditional functional alignment methods.

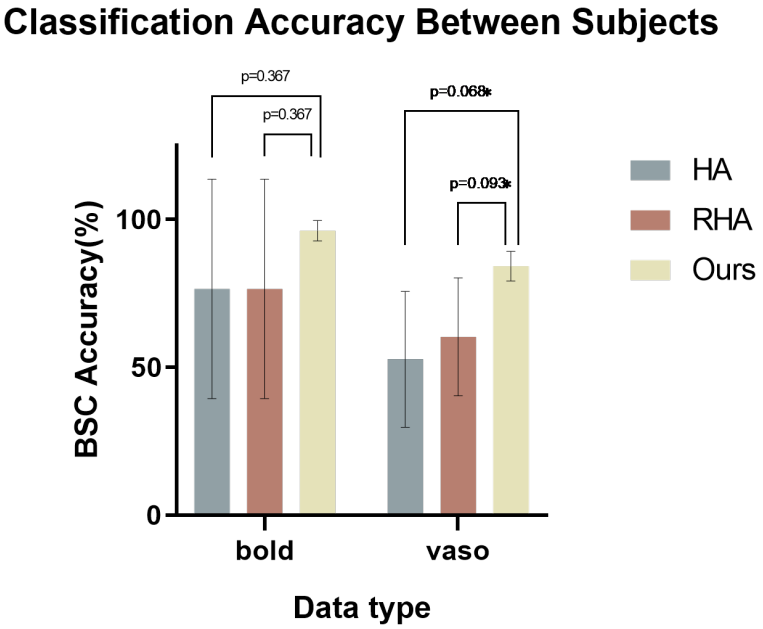

Figure 3-6 Comparison of the performance of our method with Hyper-Alignment (HA) and Regularized Hyper-Alignment (RHA) when processing new subject data on the BOLD and VASO dataset (The error bars represent standard deviation)

In addition to the two public available visual datasets mentioned above, we also conducted experiments on the 7T motion task dataset provided by Dr. Icaro Oliveira and Dr. Wietske van der Zwaag. The ROI of this dataset is selected as the primary motor cortex and includes two types of fMRI signals, BOLD and VASO, with higher resolution.

The experimental results on the 7T dataset are consistent with previous results, indicating that our method achieves higher (a trend, not statistically significant) cross subject classification accuracy on both the Bold and Vaso datasets compared to traditional functional alignment methods, with the addition of new subject data and without retraining the model.

3.5 Subject Data Migration Experiment

3.5.1 Experimental Design

In previous experiments, the experimental results have demonstrated that the method can effectively suppress the public space offset. In fact, the fundamental purpose of suppressing the common space offset is to reuse the already trained decoding model.

After adding new subject data, for those functional alignment methods that do not implement a fixed common space, it is necessary to re-train the downstream decoding model based on the new common space in order to avoid performance degradation or even failure of the decoding model. In contrast, the functional alignment method we propose can reuse the previously trained decoding models, which means that the method saves the cost of retraining the models.

In addition to saving the cost of re-training the model, using a model that has already been trained based on the original subject data is equivalent to expanding the number of training sets, as data from different subjects are mapped into the common space. We can use the aligned new subject data to continue training the decoding model. At this point, the training set includes not only data from newly added individual subjects, but also data from previous subjects, and data from previous subjects will augment the training set data and mitigate overfitting. As a result, the cross-subject generalized model will have better decoding ability than a subject-specific model trained only on new subject data. On the other hand, this also means that if we want to

achieve the same decoding performance, there will be less need for subject-specific data using the approach in this paper.

To verify that the functional alignment method in this paper achieves cross-subject data migration through cross-subject data migration and effectively reduces the amount of data required for a particular subject, the following experiments were designed. We conducted experiments on the ds000105 dataset. First, similar to the previous experiments, we use the Leave One Out (LOO) strategy to divide the dataset into the "new subject data" containing one subject and the "previous subject data" containing the rest of the subjects. In order to simulate the real situation in the application of brain-computer interface, we divide the "new subject" data into two parts. The first part of the data is called the training set, which corresponds to the training data collected in advance for training the decoding model of a particular subject in the BCI application. The other part of the data, called the test set, corresponds to the real-time brain signals generated by the subjects in the BMI application, which need to be decoded and recognized by the model. In addition, the training set of new subjects will be divided into six equal parts to test the effect of different amounts of new subject data on the decoding performance.

In order to test the effectiveness of cross-subject data migration, this paper needs to compare the performance of three different methods of utilizing new subject data on different amounts of new subject data. The first data processing method is to train a subject-specific decoding model directly using the new subject data without involving any other subject data or functional alignment methods. The second data processing method is to migrate the new subjects' data to the existing common space and then directly use the decoding models trained on all the original subjects' data for decoding. The process of data migration can only be achieved by the functional alignment method proposed in this paper. The third data processing method is to migrate the new subjects' data to the original public space and then use the migrated data to fine-tune the original decoding model. This approach entails the cost of continuing to train the decoding model but makes full use of all available data.

3.5.2 Results

In this paper, a subject-specific decoding model is first retrained using a new subject's training set, and then a test set is applied to the decoding model to test classification accuracy. The results obtained are labeled as "Train New Decoder". For the second data processing method mentioned earlier, this paper uses data from previous subjects for functional alignment

and trains a decoding model with the aligned data. Then, the paper uses the training set data of the new subject to align with the previous subject to obtain a mapping from the new subject's individual space to the public space. Next, the paper uses this mapping to transform the test set of the new subjects to the public space and eventually applies it directly to the previous decoding model. The resulting classification accuracies are labeled "Use Previous Decoder ". Corresponding to the third data processing method mentioned earlier, after aligning the new subjects with other subjects, we first convert the training set data of the new subjects to the public space, and then use the training set data of the new subjects in the public space to fine-tune the previous decoding model. Finally, we apply the test set data of the new subjects in the public space to the fine-tuned decoding model, and the classification accuracy obtained is labeled as "FineTune Previous Decoder".

We use the "leave-one-subject" method to select new subjects and repeat the experiments, taking the average value as the final performance index. In order to test the model's demand for new subjects, we also repeat the experiment on different sizes of new subjects' training datasets, and the experimental results are shown in Figure 3-7.

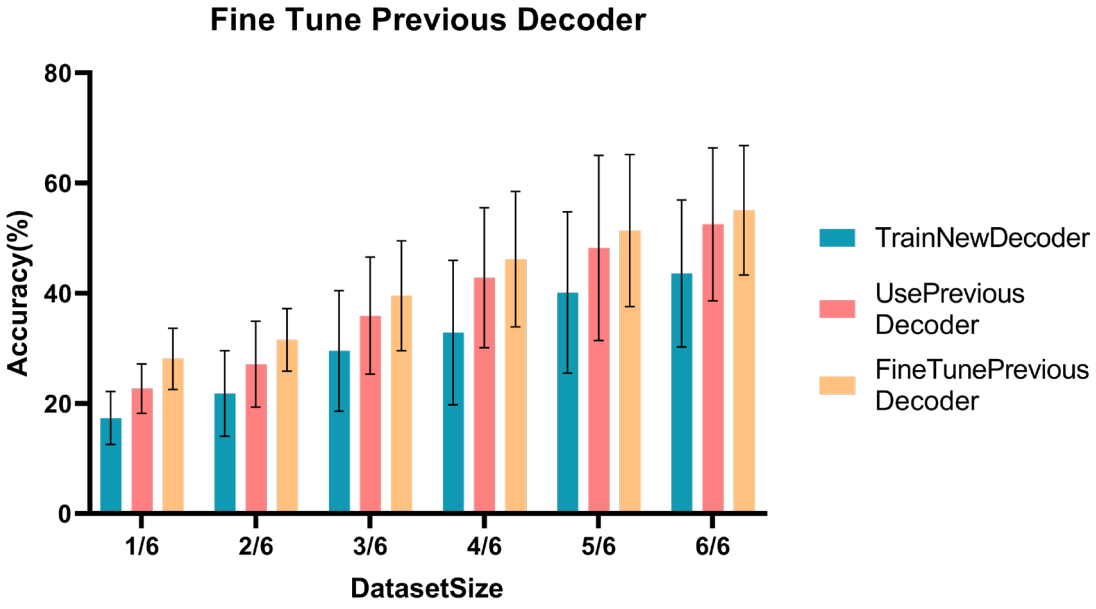

Figure 3-7 Comparison of the demand for new subject data for the three different data processing methods (The error bars represent standard deviation)

In the figure, the horizontal axis represents different numbers of new subject data and the vertical axis represents the model performance obtained using that data. From the figure, we

can find that when the same number of new subjects are used, the classification accuracy of "FineTune Previous Decoder" is higher than that of "Use Previous Decoder", while the classification accuracy of " Use Previous Decoder " is higher than that of "Train New Decoder". Since the amount of data for the previous subjects is larger than that for the new subjects, it is more efficient to use the previous model directly than to train a new subject-specific model. Spending more on training and using new subject data to further train the model increases the training set size of the decoding model, resulting in better results.

We can also see that in order to achieve the same or even better classification accuracy, the amount of new subject data required for cross-subject data migration using the method we propose is less than directly training a new subject-specific model. This demonstrates that the method in this paper can reduce the need for subject-specific data by migrating data between subjects.

3.6 Equilibrium Parameter Experiment

3.6.1 Experimental Design

Hyperparameters play a crucial role in deep learning models. Firstly, the hyperparameters are related to the optimization algorithm, which directly affect how and how fast the model parameters are updated and how convergent the training process is. The learning rate is one of the most crucial ones, which determines the step size of each parameter update. Batch size determines the number of input samples for each iteration of training, which has an impact on the stability and speed of training. Then there are the hyperparameters related to the model structure. These hyperparameters are mainly related to the architecture and fitting ability of the model. The number of hidden layer nodes determines the number of nodes in each hidden layer in the model, which directly affects the representation ability of the model. The activation function determines the output mode of neurons and affects the nonlinear fitting ability and learning effect of the model. In order to exclude the influence of hyperparameter selection in the experimental results as much as possible, in all the above experiments we adopt completely fixed training hyperparameters. In terms of model structure hyperparameters, we use a 4-layer neural network for both encoder and decoder. For the encoder, the number of neurons in each layer is (4096, 2048, 4096, 2048), and for the decoder, the number of neurons in each layer is (2048, 4096, 2048, 4096), where a nonlinear activation layer with the ReLu function as the activation function is included between each layer to achieve nonlinear mapping.

For the approach in this paper, the model hyperparameters come from the equilibrium parameters in the loss function in addition to the model structure λ .

For the hyperparameter in the loss function λ , the relevant formula is shown in Eq. 3-6:

$$\begin{aligned} Loss\ 1 &= \left\| x - x_{gen} \right\|_2 \\ Loss\ 2 &= \left\| v_{pre} - v_{gt} \right\|_2 \\ Loss &= \lambda Loss1 + (1-\lambda) Loss2 \end{aligned} \tag{3-6}$$

Where $Loss1$ is the reconstruction loss of the self-encoder, and $Loss\ 2$ is the contrast learning loss on the common space. The hyperparameters $\boldsymbol{\lambda}$ is the balance parameter.

For loss function design based on contrast learning, we use generalized constraints rather than pairwise constraints. Pairwise constraints refer to explicit constraints on the pairwise relationships between data in contrast learning, while generalized constraints use broader constraints to guide the model in learning the relationships between data. Pairwise constraints tend to require a large amount of pairwise labeled data, which is detrimental to relatively scarce neural signals. In addition to this, pairwise constraints also tend to be difficult to adapt to new or unseen pairwise relationships, which would also be detrimental to cross-subject data migration. Therefore, we design the loss function as well as the model based on generalized constraints to enhance the flexibility and generality of the model and to reduce the acquisition requirements for fMRI signals.

In the previous experiments, in order to exclude the interference of hyperparameter λ selection on the experimental results, we uniformly used 0.5 as the default λ value. However, in order to assess the impact of the balance parameter in the loss function on the model alignment performance and to verify the effectiveness of the comparison function we selected in the model optimisation process, we conduct relevant experiments on the ds000105 dataset.

The specific experimental procedure is as follows:

a. Select a λ value.

b. Align the data of all N subjects to the common space using a loss function based on this λ value of the loss function, respectively, to align the data of all N subjects to the common space.

c. The leave-one-out method selects one subject.

d. Train an SVM classifier for the classification task on the remaining aligned N-1 subject data.

e. The post-alignment data of the subject selected by the leave-one-out method was used as the test set.

f. Assess cross-subject classification accuracy.

g. Repeat steps c-f in a cross-validation manner to obtain an average accuracy.

h. Count the accuracy of the alignment model based on this λ value and repeat the process to calculate the next λ value.

3.6.2 Results

The effect of balancing Parameters λ on alignment performance is shown in Figure 3-8.

Between the proportion of balanced weights from 0 to 1, the larger the value taken, the smaller the corresponding proportion of comparative learning loss in the loss function. From the figure, we can find that on the DS105 dataset when λ is smaller than 0.7, the decoding performance can be maintained at a stable level, and when λ greater than 0.7, the decoding performance will show a significant decrease. This suggests that increasing and maintaining a certain weight of contrast learning loss will be beneficial to achieve better alignment performance. This result demonstrates the effectiveness of the contrast learning loss chosen as the key objective function of the model, which makes the brain activity patterns for similar stimuli or tasks as close as possible and as far away as possible from the brain activity patterns for different stimuli or tasks between subjects.

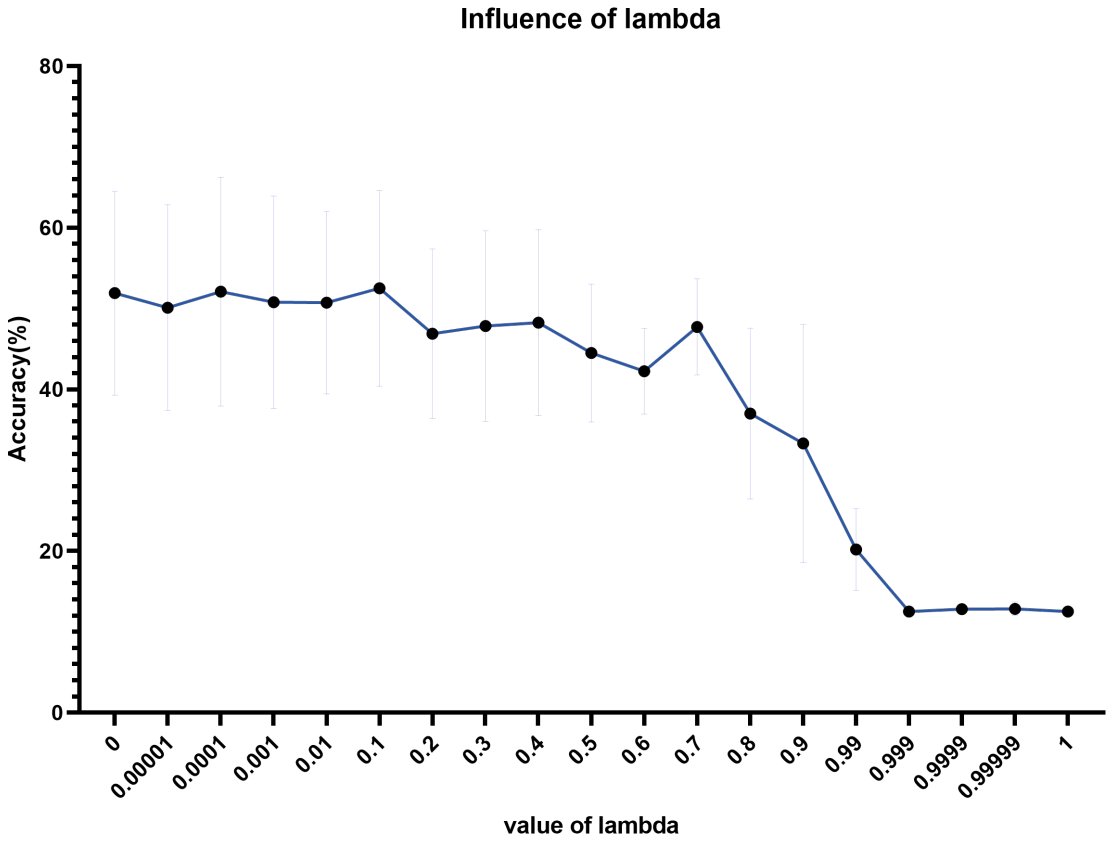

Figure 3-8 Balancing parameters in the loss function λ effect on alignment performance (The error bars represent standard deviation)

3.7 Summary and Discussion

The time and economic cost of brain signal acquisition is usually very expensive for individual subjects, which will bring about an insufficient amount of subject data. Meanwhile, decoding models for brain-computer interfaces generally suffer from data starvation, which will lead to overfitting of decoding models, thus making the decoding performance degraded. In order to solve the data problem that limits the development and application of brain-computer interfaces to a certain extent, we propose to implement a cross-subject data migration method, which migrates the data of other subjects to a common feature space, in order to obtain a large amount of cross-subject common data for the common decoding model, without having to make each subject go through a long process of brain signal acquisition. Therefore, this paper proposes a feature alignment model suitable for subject data migration to solve this problem. The feature alignment model is based on a set of extensible self-encoder frameworks that can align arbitrary newly added subject data to a fixed common feature space without introducing other subject information at all.

The experimental results show that for fMRI data, the method we propose has a higher cross-subject classification accuracy on the newly added subject data experiments compared to both traditional functional alignment methods represented by hyper-alignment. This indicates that our proposed method successfully implements a fixed common space for the alignment process, effectively mitigating the performance degradation due to the spatial transformations and allowing the new subject data to be used on the previously trained model without the need to retrain the downstream decoding model. At the same time, the scalability of the framework proposed in this paper in terms of new-subject alignment also allows us to use the new-subject data to fine-tune the models trained on other-subject data. Fine-tuning the downstream decoding model after aligning the new-subject data to the common space utilizes the data from other subjects, which is equivalent to migrating the other-subject data to the new-subject decoding model, thus expanding the number of training sets for the downstream decoding model. The experimental results show that with a limited amount of new subject data, aligning the new subject data using this paper's method yields better classification accuracy than re-training a new model with the new subject data, which proves that this paper's method successfully achieves cross-subject data migration, and demonstrates the effectiveness of the proposed method in lowering the cost of data acquisition with the potential of solving the problem of insufficient amount of data for the brain-computer interface.

In terms of model structure design and computational process design, the method we propose communicates the alignment process between each group of self-encoders and each subject through a common orthogonal matrix, decouples the alignment process of different subjects in time and space, realizes asynchronous and distributed alignment, and improves the processing efficiency and scalability of alignment. With the continuous increase of the number of subjects, the time complexity of the traditional functional alignment method grows exponentially, while the time complexity of this paper's method grows linearly, so the computation time required by this paper's method grows relatively slowly, which is more suitable for processing large-scale data. The traditional functional alignment method requires linearly increasing memory, while the method in this paper only requires a fixed size of memory due to the decoupling of the computational space, which reduces the requirement for computation devices when processing large-scale data; for new subject data, the method in this paper also does not need to introduce other subjects' data for alignment, which further reduces the complexity of the whole alignment process and makes it more flexible and efficient when

processing new subject data. Therefore, the method in this paper has obvious advantages and broad prospects in dealing with the ever-growing subject data.

In previous functional alignment studies, the optimization of the alignment model relied on the neural synchrony resulting from each subject receiving the same stimulus or performing the same task, a one-to-one mapping relationship that relies on a large amount of precisely labeled data. By exploiting the generalized constraints of contrast learning rather than pairwise constraints, the alignment model in this paper does not rely on fMRI data that is perfectly synchronized with the inter-subject stimuli, which further reduces the requirement for neural signal data collection.

4. Conclusion and Discussion

4.1 Conclusion

In this paper, we propose corresponding solutions to two problems facing cross-subject generalized neural decoding.

Firstly, in Chapter 2, this paper proposes DecodeGPT, a universal visual decoding model based on GPT, to address the problem that each subject and each task needs to train an ad hoc decoding model. The innovation of this method is: firstly, it establishes a multimodal language model that understands the brain signals and the text, and implements the human language text control using the prompt-tune strategy to achieve cross-subject decoding behavior; secondly, the method also makes full use of whole-brain information through the mechanism of multi-head cross-attention, and in this way improves the overall decoding performance of the model. DecodeGPT is able to achieve multi-task and cross-subject decoding within a single model, and has excellent decoding performance on each task per subject. In order to achieve multi-subject and cross-subject decoding, we designed Prompt text instructions containing "subject + task" information, and fine-tuned GPT-2 based on the Prompt text instructions and the corresponding fMRI sample dataset. In this way, GPT-2 is able to process the potential features after adding the brain signal features, and decode the corresponding subjects and tasks. In the classification task, the model generates predicted labeled words directly through the GPT text generator. Although the decoding results of the classification model based on the GPT language model are derived from the generated text of the language model, the chance level is therefore close to 0 from a probabilistic point of view. Nevertheless, the model in still exhibits a good classification performance on the classification task for each subject, well above the chance level of 12 classifications. In the confusion matrix, the approach shows a greater probability of confusion for semantically similar categories, such as "kitchen" vs "food", "appliances" and "electronics". This suggests that the model successfully establishes a meaningful and generalized mapping between each subject's brain visual stimulus response patterns and semantic concepts. In the caption decoding task, the model generates natural language text descriptions that can accurately reflect the content of the original visual stimulus images, and there is a high semantic similarity between the generated text descriptions and the descriptions from human language, far beyond the random level. Compared with some current works that perform caption decoding, the method has better overall performance on the caption decoding

task, which may be brought about by the GPT language model obtained based on a large amount of corpus pre-training.

In Chapter 3, we propose a functionally aligned model for subject data migration to address the problem of universal model retraining caused by the growing number of subjects. The innovation of this approach is the creation of an easily extensible functional alignment model architecture, which allows for better performance and lower computational time-space overhead in dealing with the increasing number of new subject data; in addition to this, the approach also reduces the need for neural signal data acquisition through the application of generalized contrast learning. The feature alignment model is based on a set of extensible self-encoder frameworks that can align any newly added subject data to a fixed common feature space without introducing additional subject information. The method successfully implements a fixed common space, mitigating the performance degradation due to spatial transformations, allowing new subject data to be used on previously trained models without the need to retrain downstream decoding models. The method also implements cross-subject data migration, migrating data from other subjects to a common feature space, providing a large amount of cross-subject common data for a universal decoding model without the need for an extensive brain signal acquisition process. By communicating the groups of self-encoders through a common orthogonal matrix, the method has an asynchronous and distributed nature in the alignment process, which improves the processing efficiency and scalability of the alignment. Compared with traditional methods, the time complexity of the method increases linearly, which is more suitable for processing large-scale data; moreover, the method does not need to introduce other subjects' data for alignment, which further reduces the complexity when processing new subjects' data. In addition, the method does not rely on stimulus-synchronized fMRI data, which reduces the data collection requirements by exploiting the generalized constraints of contrast learning.

In summary, this paper develops the current cross-subject universal decoding method from two directions, namely multi-tasking and cross-subject data migration, respectively, and expands the applicability of the universal decoding model in the case of multi-tasking and growth in the number of subjects, which provides assistance in the development of brain-computer interface models with more practical applications.

4.2 Discussion

In this thesis, we propose a universal decoding method for multi-tasking and cross-subject data migration, and conduct a series of experiments to validate the performance of the decoding method, but there are still some areas for improvement and enhancement during the experiments:

1. There is an imbalance in the number of samples with different labels in the visual stimulus dataset of NSD, and the labels themselves do not uniquely reflect all the semantic information of a natural image, so the quality of the dataset may limit the decoding model from accurately learning the mapping relationship between brain signals and semantic information.
2. Furthermore, due to the size constraints of the GPT model used, the representation of brain signals is limited by relatively low dimensionality, which may affect the ability of the decoding model in this paper to decode more detailed and accurate semantic information.
3. The number of subjects in the dataset used is still small, making it difficult to directly measure the performance of the model in various aspects under a massive number of subjects.
4. Despite the inclusion of different types of fMRI signals, the type of neural signals used for the experiments remains relatively homogeneous, and theoretically, the method proposed in this paper can be applied to all forms of neural signals including EEG, MEG, and NIRS, etc., and thus subsequent experiments can be performed on a wider range of forms of neural signals.
5. For the multi-task cross-subject universal decoding model, since the GPT-2 model itself does not have multimodal properties, and since our approach does not establish a mapping of brain signals to textual information, the ability to understand multimodality in our approach comes entirely from the fine-tuning phase of the training process. Due to the limited nature of the multimodal samples in prompt-tune, the model is only able to recognize the textual instructions used in prompt-tune, which somewhat limits the flexibility of the model to understand the information. If this is to be improved, multimodal large language models that already have image understanding capabilities can be used as the base model, such as Minigpt-4 (Zhu, Chen, Shen, Li, & Elhoseiny, 2023) , LLaVA(H. Liu, Li, Wu, & Lee, 2024) etc., and

migrate the model's image comprehension capabilities to brain signal understanding by means of cross-modal alignment. There has been work in previous research that has achieved the alignment of image features with brain signals, for example, Liu et al. used CLIP image text representations to guide brain signal representations, and constructed BrainCilp to achieve the tri-modal alignment of brain signals, linguistic text and natural image representations (Y. Liu, Ma, Zhou, Zhu, & Zheng, 2023a) The BraVL proposed by Du et al. similarly achieves this through hybrid expert modelling (C. Du, Fu, Li, He, & Intelligence, 2023) .

6. For universal decoding models oriented towards subject data migration, the actual computational complexity will be higher than traditional functional alignment methods based on typical correlation analysis when the number of subjects is fixed or very small, due to the inherent complexity of the self-encoder set used. In addition to this, relying solely on the encoder loss in the loss function to retain sample-specific information may lead to a reduction in the ability of the downstream decoding model to decode detailed information in cases where the value of the balancing parameter in the loss function takes a very small value.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., . . . Anadkat, S. J. a. p. a. (2023). Gpt-4 technical report.

Akaho, S. J. a. p. c. (2006). A kernel method for canonical correlation analysis.

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., . . . Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat Neurosci, 25*(1), 116-126. doi:10.1038/s41593-021-00962-x

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., . . . Chen, Z. J. a. p. a. (2023). Palm 2 technical report.

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature, 568*(7753), 493-498. doi:10.1038/s41586-019-1119-1

Banerjee, S., & Lavie, A. (2005). *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments.* Paper presented at the Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.

Benton, A., Khayrallah, H., Gujral, B., Reisinger, D. A., Zhang, S., & Arora, R. J. a. p. a. (2017). Deep generalized canonical correlation analysis.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. J. T. o. t. a. f. c. l. (2017). Enriching word vectors with subword information. *5*, 135-146.

Bressler, S. L., & Menon, V. (2010). Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn Sci, 14*(6), 277-290. doi:10.1016/j.tics.2010.04.004

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . Askell, A. J. A. i. n. i. p. s. (2020). Language models are few-shot learners. *33*, 1877-1901.

Buzsáki, G., Anastassiou, C. A., & Koch, C. J. N. r. n. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *13*(6), 407-420.

Chatterjee, S., & Samanta, D. J. a. p. a. (2023). DreamCatcher: Revealing the Language of the Brain with fMRI using GPT Embedding.

Chen, M., Han, J., Hu, X., Jiang, X., Guo, L., & Liu, T. (2014). Survey of encoding and decoding of visual stimulus via FMRI: an image analysis perspective. *Brain Imaging Behav, 8*(1), 7-23. doi:10.1007/s11682-013-9238-z

Chen, P.-H. C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. J. A. i. n. i. p. s. (2015). A reduced-dimension fMRI shared response model. *28*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. J. a. p. a. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., . . . Gehrmann, S. J. J. o. M. L. R. (2023). Palm: Scaling language modeling with pathways. *24*(240), 1-113.

Collinger, J. L., Wodlinger, B., Downey, J. E., Wang, W., Tyler-Kabara, E. C., Weber, D. J., . . . Schwartz, A. B. (2013). High-performance neuroprosthetic control by an individual with tetraplegia. *Lancet, 381*(9866), 557-564. doi:10.1016/S0140-6736(12)61816-9

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. J. a. p. a. (2022). Semantic scene descriptions as an objective of human vision.

Du, C., Fu, K., Li, J., He, H. J. I. T. o. P. A., & Intelligence, M. (2023). Decoding visual neural representations by multimodal learning of brain-visual-linguistic features.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. J. a. p. a. (2021). Glm: General language model pretraining with autoregressive blank infilling.

Fang, T., Qi, Y., & Pan, G. J. A. i. N. I. P. S. (2020). Reconstructing perceptive images from brain activity by shape-semantic GAN. *33*, 13038-13048.

Ferrante, M., Boccato, T., & Toschi, N. J. a. p. a. (2023). Through their eyes: multi-subject Brain Decoding with simple alignment techniques.

Ferrante, M., Ozcelik, F., Boccato, T., VanRullen, R., & Toschi, N. J. a. p. a. (2023). Brain Captioning: Decoding human brain activity into images and text.

Ferrari, M., & Quaresima, V. J. N. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *63*(2), 921-935.

Fisher, R. A. J. A. o. e. (1936). The use of multiple measurements in taxonomic problems. *7*(2), 179-188.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., . . . Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature, 536*(7615), 171-178. doi:10.1038/nature18933

Graves, A., & Graves, A. J. S. s. l. w. r. n. n. (2012). Long short-term memory. 37-45.

Han, K., Wen, H., Shi, J., Lu, K. H., Zhang, Y., Fu, D., & Liu, Z. (2019). Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *Neuroimage, 198*, 125-136. doi:10.1016/j.neuroimage.2019.05.039

Hanson, S. J., Matsuka, T., & Haxby, J. V. J. N. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *, 23*(1), 156-166.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430. doi:10.1126/science.1063736

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., . . . Ramadge, P. J. J. N. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *72*(2), 404-416.

Horikawa, T., & Kamitani, Y. J. N. c. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *8*(1), 15037.

Horikawa, T., Tamaki, M., Miyawaki, Y., & Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science, 340*(6132), 639-642. doi:10.1126/science.1234330

Hu, P., Peng, X., Zhu, H., Zhen, L., Lin, J., Yan, H., & Peng, D. J. I. T. o. C. (2021). Deep semisupervised multiview learning with increasing views. *52*(12), 12954-12965.

Huang, W., Yan, H., Cheng, K., Wang, C., Li, J., Wang, Y., . . . Zuo, Z. J. N. N. (2021). A neural decoding algorithm that generates language from visual activity evoked by natural images. *144*, 90-100.

Huang, W., Yan, H., Cheng, K., Wang, Y., Wang, C., Li, J., . . . Chen, H. J. H. B. M. (2021). A dual‐channel language decoding from brain activity with progressive transfer training. *42*(15), 5089-5100.

Huang, W., Yan, H., Wang, C., Li, J., Yang, X., Li, L., . . . Chen, H. (2020). Long short-term memory-based neural decoding of object categories evoked by natural images. *Hum Brain Mapp, 41*(15), 4442-4453. doi:10.1002/hbm.25136

Huang, W., Yan, H., Wang, C., Li, J., Zuo, Z., Zhang, J., . . . Chen, H. J. A. o. B. E. (2020). Perception-to-image: Reconstructing natural images from the brain activity of visual perception. *48*, 2323-2332.

Huang, W., Yan, H., Wang, C., Yang, X., Li, J., Zuo, Z., . . . Chen, H. J. N. b. (2021). Deep natural image reconstruction from human brain activity based on conditional progressively growing generative adversarial networks. *37*, 369-379.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. J. N. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *76*(6), 1210-1224.

Kaiser, D., Azzalini, D. C., & Peelen, M. V. (2016). Shape-independent object category responses revealed by MEG and fMRI decoding. *J Neurophysiol, 115*(4), 2246-2250. doi:10.1152/jn.01074.2015

Khademi, Z., Ebrahimi, F., & Kordy, H. M. J. J. o. N. M. (2023). A review of critical challenges in MI-BCI: From conventional to deep learning methods. *383*, 109736.

Li, W., Liu, M., Chen, F., & Zhang, D. (2020). *Graph-based decoding model for functional alignment of unaligned fMRI data.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Lin, C.-Y. (2004). *Rouge: A package for automatic evaluation of summaries.* Paper presented at the Text summarization branches out.

Lin, S., Sprague, T., & Singh, A. K. J. a. p. a. (2022). Mind Reader: Reconstructing complex images from brain activities.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). *Microsoft coco: Common objects in context.* Paper presented at the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13.

Lin, Y., Li, J., & Wang, H. (2019). *Dcnn-gan: Reconstructing realistic image from fmri.* Paper presented at the 2019 16th International Conference on Machine Vision Applications (MVA).

Liu, H., Li, C., Wu, Q., & Lee, Y. J. J. A. i. n. i. p. s. (2024). Visual instruction tuning. *36*.

Liu, Y., Ma, Y., Zhou, W., Zhu, G., & Zheng, N. J. a. p. a. (2023a). BrainCLIP: Bridging Brain and Visual-Linguistic Representation Via CLIP for Generic Natural Visual Stimulus Decoding.

Liu, Y., Ma, Y., Zhou, W., Zhu, G., & Zheng, N. J. a. p. a. (2023b). BrainCLIP: Bridging Brain and Visual-Linguistic Representation via CLIP for Generic Natural Visual Stimulus Decoding from fMRI.

Logothetis, N. K. J. N. (2008). What we can do and what we cannot do with fMRI. *453*(7197), 869-878.

Lorbert, A., & Ramadge, P. J. J. A. i. N. I. P. S. (2012). Kernel hyperalignment. *25*.

Luo, A. F., Henderson, M. M., Tarr, M. J., & Wehbe, L. J. a. p. a. (2023). BrainSCUBA: Fine-Grained Natural Language Captions of Visual Cortex Selectivity.

Mai, W., & Zhang, Z. J. a. p. a. (2023). Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity.

Matsuo, E., Kobayashi, I., Nishimoto, S., Nishida, S., & Asoh, H. (2018). *Describing semantic representations of brain activity evoked by visual stimuli.* Paper presented at the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. J. a. p. a. (2013). Efficient estimation of word representations in vector space.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M. A., Morito, Y., Tanabe, H. C., . . . Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron, 60*(5), 915-929. doi:10.1016/j.neuron.2008.11.004

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron, 63*(6), 902-915. doi:10.1016/j.neuron.2009.09.006

Nishida, S., & Nishimoto, S. J. N. (2018). Decoding naturalistic experiences from human brain activity via distributed representations of words. *180*, 232-242.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Curr Biol, 21*(19), 1641-1646. doi:10.1016/j.cub.2011.08.031

O'toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. J. J. o. c. n. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *17*(4), 580-590.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., . . . Ray, A. J. A. i. n. i. p. s. (2022). Training language models to follow instructions with human feedback. *35*, 27730-27744.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *Bleu: a method for automatic evaluation of machine translation.* Paper presented at the Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

Pennington, J., Socher, R., & Manning, C. D. (2014). *Glove: Global vectors for word representation.* Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

Pina, L., Sien, S.-W., Song, C., Ward, T. M., Fogarty, J., Munson, S. A., & Kientz, J. A. J. P. o. t. A. o. H.-c. I. (2020). DreamCatcher: exploring how parents and school-age children can track and review sleep information together. *4*(CSCW1), 1-25.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., . . . Clark, J. (2021). *Learning transferable visual models from natural language supervision.* Paper presented at the International conference on machine learning.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. J. O. b. (2019). Language models are unsupervised multitask learners. *1*(8), 9.

Rakhimberdina, Z., Jodelet, Q., Liu, X., & Murata, T. (2021). Natural Image Reconstruction From fMRI Using Deep Learning: A Survey. *Front Neurosci, 15*, 795488. doi:10.3389/fnins.2021.795488

Schick, T., & Schütze, H. J. a. p. a. (2020). Exploiting cloze questions for few shot text classification and natural language inference.

Serruya, M. D., Hatsopoulos, N. G., Paninski, L., Fellows, M. R., & Donoghue, J. P. (2002). Instant neural control of a movement signal. *Nature, 416*(6877), 141-142. doi:10.1038/416141a

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. J. P. c. b. (2019). Deep image reconstruction from human brain activity. *15*(1), e1006633.

Shin, H. C., Aggarwal, V., Acharya, S., Schieber, M. H., & Thakor, N. V. (2010). Neural decoding of finger movements using Skellam-based maximum-likelihood decoding. *IEEE Trans Biomed Eng, 57*(3), 754-760. doi:10.1109/TBME.2009.2020791

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., . . . Neal, D. J. a. p. a. (2023). Towards expert-level medical question answering with large language models.

Song, S., Zhan, Z., Long, Z., Zhang, J., & Yao, L. (2011). Comparative study of SVM methods combined with voxel selection for object category classification on fMRI data. *PLoS One, 6*(2), e17191. doi:10.1371/journal.pone.0017191

Sun, S., Xie, X., & Yang, M. (2016). Multiview Uncorrelated Discriminant Analysis. *IEEE Trans Cybern, 46*(12), 3272-3284. doi:10.1109/TCYB.2015.2502248

Takada, S., Togo, R., Ogawa, T., & Haseyama, M. (2020). *Generation of viewed image captions from human brain activity via unsupervised text latent space.* Paper presented at the 2020 IEEE International Conference on Image Processing (ICIP).

Takagi, Y., & Nishimoto, S. (2023). *High-resolution image reconstruction with latent diffusion models from human brain activity.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Teng, C., & Kravitz, D. J. (2019). Visual working memory directly alters perception. *Nat Hum Behav, 3*(8), 827-836. doi:10.1038/s41562-019-0640-4

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., . . . Azhar, F. J. a. p. a. (2023). Llama: Open and efficient foundation language models.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., . . . Bhosale, S. J. a. p. a. (2023). Llama 2: Open foundation and fine-tuned chat models.

VanRullen, R., & Reddy, L. (2019). Reconstructing faces from fMRI patterns using deep generative neural networks. *Commun Biol, 2*(1), 193. doi:10.1038/s42003-019-0438-y

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. J. A. i. n. i. p. s. (2017). Attention is all you need. *30*.

Wakeman, D. G., & Henson, R. N. J. S. d. (2015). A multi-subject, multi-modal human neuroimaging dataset. *2*(1), 1-10.

Wen, H., Shi, J., Zhang, Y., Lu, K. H., Cao, J., & Liu, Z. (2018). Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cereb Cortex, 28*(12), 4136-4160. doi:10.1093/cercor/bhx268

Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., & Shenoy, K. V. High-performance brain-to-text communication via imagined handwriting 1 2.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., . . . Mann, G. J. a. p. a. (2023). Bloomberggpt: A large language model for finance.

Xu, H., Lorbert, A., Ramadge, P. J., Guntupalli, J. S., & Haxby, J. V. (2012). *Regularized hyperalignment of multi-set fMRI data.* Paper presented at the 2012 IEEE statistical signal processing workshop (SSP).

Yousefnezhad, M., Selvitella, A., Han, L., Zhang, D. J. I. T. o. C., & Systems, D. (2020). Supervised hyperalignment for multisubject fmri data alignment. *13*(3), 475-490.

Yousefnezhad, M., & Zhang, D. J. A. i. N. I. P. S. (2017). Deep hyperalignment. *30*.

Zaremba, W., Sutskever, I., & Vinyals, O. J. a. p. a. (2014). Recurrent neural network regularization.

Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. J. a. p. a. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models.