Wills Aligner: Multi-Subject Collaborative Brain Visual Decoding

Guangyin Bao[1], Qi Zhang[1], Zixuan Gong[1], Jialei Zhou[1],
Wei Fan[2], Kun Yi[3, 4], Usman Naseem[5], Liang Hu[1], Duoqian Miao[1*]
[1]Tongji University
[2]University of Oxford
[3]North China Institute of Computing Technology
[4]State Information Center of China
[5]Macquarie University
{baogy, zhangqi_cs, gongzx, dqmiao}@tongji.edu.cn

Abstract

Decoding visual information from human brain activity has seen remarkable advancements in recent research. However, the diversity in cortical parcellation and fMRI patterns across individuals has prompted the development of deep learning models tailored to each subject. The personalization limits the broader applicability of brain visual decoding in real-world scenarios. To address this issue, we introduce Wills Aligner, a novel approach designed to achieve multi-subject collaborative brain visual decoding. Wills Aligner begins by aligning the fMRI data from different subjects at the anatomical level. It then employs delicate mixture-of-brain-expert adapters and a meta-learning strategy to account for individual fMRI pattern differences. Additionally, Wills Aligner leverages the semantic relation of visual stimuli to guide the learning of inter-subject commonality, enabling visual decoding for each subject to draw insights from other subjects' data. We rigorously evaluate our Wills Aligner across various visual decoding tasks, including classification, cross-modal retrieval, and image reconstruction. The experimental results demonstrate that Wills Aligner achieves promising performance.

Introduction

The process of human perception is marvelous. Our perception of the world is shaped not only by the objective reality around us but also by our individual subjective experiences. Understanding the mechanisms of human perception is crucial for unraveling the complexities of the brain, advancing brain-inspired computational models (Du et al. 2023; Palazzo et al. 2020; Pereira et al. 2018; Nie et al. 2023), and offering numerous applications in clinical medicine (Linden 2021) and brain-computer interfaces (Naselaris et al. 2011; Kamitani and Tong 2005). Within this expansive field, visual decoding stands out as a critical and challenging study. It enables us to delve into the intricate workings of the brain during visual processing, object recognition, and scene interpretation (Parthasarathy et al. 2017; Zhang et al. 2022). Among the various brain imaging modalities, functional magnetic resonance imaging (fMRI) is particularly favored by researchers due to its non-invasive nature and its ability to precisely localize functional regions of the cortex. Consequently, fMRI-based visual decoding has become a significant and prominent topic in neuroscience research.

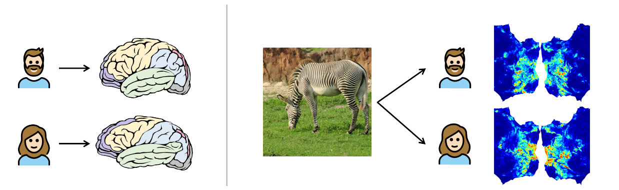


Figure 1: Illustrations of inter-subject brain differences. The left figure illustrates structural differences, showing variations in size and location of the same functional areas (highlighted in the same color) across subjects. The right figure depicts differences in fMRI patterns, where distinct brain activities are observed when the same stimulus is presented.

Extensive studies (Horikawa and Kamitani 2017; Lin, Sprague, and Singh 2022; Ozcelik and VanRullen 2023; Scotti et al. 2023; Liu et al. 2023) have investigated brain visual decoding. They construct classification, retrieval, and image reconstruction tasks to mine visual information from fMRI. However, these methods encounter significant challenges in practical application, primarily because they are customized for individual subjects. To elaborate, a subject-specific deep model is developed by utilizing the subject's fMRI data, necessitating an expanding number of models due to the substantial differences in brains. These differences arise from variations in brain structure and fMRI patterns, attributed to diverse genetic backgrounds and cognitive development processes. As shown in Figure 1, structural differences pertain to variations in cortical parcellation among subjects, with disparities in the size and location of functional areas like visual, language, and memory regions. It can only be identified through anatomical information such as gray and white matter (Fischl et al. 1999; Destrieux et al. 2010). The fMRI pattern differences refer to the fact that different subjects exhibit distinct brain activity even when exposed to the same visual stimulus. These factors complicate the development of multi-subject visual decoding models. Hence, it is crucial to explore feasbile approaches that can achieve effective collaborative visual decoding across subjects while accommodating individual differences.

[*]Corresponding author.

Currently, research on multi-subject collaboration in visual decoding is quite limited, and existing multi-subject models often underperform compared to single-subject models. Addressing structural differences between subjects is a prerequisite for effective collaboration. Recent methods (Zhou et al. 2024; Wang et al. 2024) tackle this challenge with basic data preprocessing techniques like PCA or max pooling, which, however, can lead to brain information loss and misalignment of functional regions. Other approaches (Ferrante, Boccato, and Toschi 2023) achieve fMRI structural alignment by training linear models based on the same images viewed by different subjects. These methods also struggle with scalability due to the need for identical visual stimuli across all subjects. Fortunately, advances in neuroscience offer inspiration by leveraging prior anatomical knowledge and aligning fMRI data to a standardized brain template. Besides achieving fMRI alignment, another challenge is to perform effective multi-subject collaboration. Existing approaches either employ subject-specific tokens to identify each subject (Zhou et al. 2024) or adapt the initial ridge regression to individual subjects (Scotti et al. 2024), allowing model training to accommodate individual differences. Although making desirable progress, these methods only achieve superficial subject information integration from the data input aspect rather than deep semantics of representations. In addition, they have primarily focused on handling fMRI pattern differences while overlooking inter-subject commonality, resulting in visual decoding that benefits only from data within a single subject rather than multiple subjects. Naturally, it is necessary to learn both the inter-subject commonality and perceive various fMRI patterns for achieving multi-subject collaboration.

In light of the above discussion, we propose Wills Aligner. We employ an fMRI alignment technique derived from neuroscience, termed anatomical alignment, to address brain structural differences. To facilitate effective multi-subject collaboration, we first employ the semantic relation of visual stimuli to guide the model in learning inter-subject commonality, thereby enabling the transfer of universal fMRI visual decoding knowledge among subjects. Following this, we introduce Mixture-of-Brain-Expert (MoBE) adapters, which are subject-guided sparse MoE networks designed to capture distinct fMRI patterns. Additionally, we implement a meta-learning strategy that progressively integrates these learned fMRI patterns into the semantics of deep representations, enhancing multi-subject decoding performance.

We conducted a comprehensive evaluation on the Natural Scene Dataset (NSD) (Allen et al. 2022), including multi-label classification, bidirectional retrieval, few-shot learning, and fMRI-to-image reconstruction. Our experimental results demonstrate that the Wills Aligner consistently delivers promising performance, surpassing both single-subject and multi-subject baselines under equivalent conditions.

Our contributions are summarized as follows:

- We identify key limitations in existing visual decoding methods and propose a multi-subject collaborative approach that addresses these challenges.
- We utilize the anatomical alignment to address structural differences in fMRI data across subjects, experimentally demonstrating its superiority over alternative fMRI preprocessing methods.
- We leverage the semantic relation of visual stimuli to guide inter-subject commonality learning and introduce MoBE adapters with a meta-learning strategy to capture different fMRI patterns.

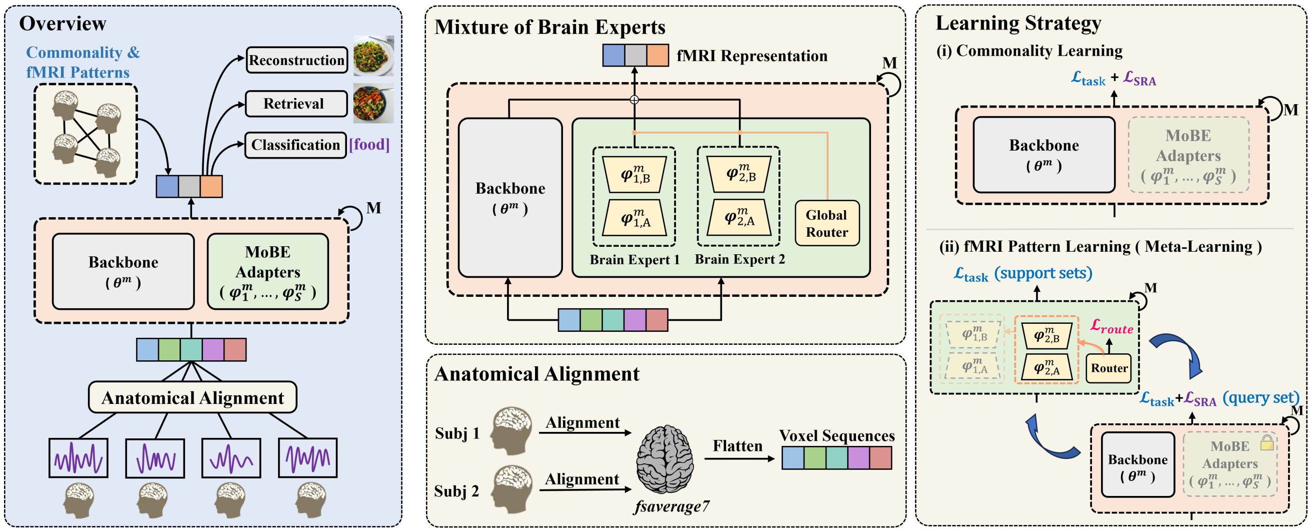
Related Works

fMRI Visual Decoding

The study of brain visual decoding based on fMRI has been a long-standing endeavor (Lu et al. 2023; Scotti et al. 2023, 2024; Gong et al. 2024b,a; Huo et al. 2025; Li et al. 2025). Due to the low signal-to-noise ratio of fMRI, early studies (Takagi and Nishimoto 2023) emphasized the use of purely linear models, such as linear regression, to embed fMRI and images into a common intermediate space. Further studies (Du et al. 2023; Horikawa and Kamitani 2017; Shen et al. 2019; Lin, Sprague, and Singh 2022) used VGG or ResNet to extract semantically rich image representations. Mindeye (Scotti et al. 2023) demonstrated that a large MLP can serve as an excellent brain representation learner. Other work (Zhou et al. 2024; Chen et al. 2023; Chen, Qing, and Zhou 2023; Gong et al. 2024c) attempted to use Transformer-style (Vaswani et al. 2017) in brain visual decoding. As for optimization objectives, early studies (Horikawa and Kamitani 2017; Roelfsema, Denys, and Klink 2018) used the cross-entropy of categorization tasks as supervised objectives. Recognizing visual stimuli categories from fMRI can decode coarse-grained visual information. Mind-Reader (Lin, Sprague, and Singh 2022) first used an fMRI-to-image retrieval task to decode fine-grained visual information. Inspired by contrastive learning (Radford et al. 2021), Brain-Diffuser (Ozcelik and VanRullen 2023) aligned fMRI representations to pre-trained ViT's latent space. MindEye (Scotti et al. 2023) proposed BiMixCo, a data enhancement strategy, to achieve high-performance fMRI-to-image reconstruction. Some other work explored self-supervised objectives, such as MinD-Vis (Chen et al. 2023) and MindVideo (Chen, Qing, and Zhou 2023) used masked brain modeling. However, Due to the large variation among subjects, the above methods only trained proprietary single-subject models for each subject.

Multi-Subject Visual Decoding

Some advanced studies have explored multi-subject brain visual decoding. They focus on how to align fMRI data from various subjects and how to learn inter-subject differences. CLIP-MUSED (Zhou et al. 2024) first studies multi-subject visual classification, and its performance reaches the level of a single-subject model. MindBridge (Wang et al. 2024) aligns fMRI using adaptive fMRI aggregation. Psychometry (Quan et al. 2024) learns inter-subject differences using dense MoE with subject-specific projector layers. MindEye2 (Scotti et al. 2024) aligns fMRI using simple subject-specific ridge regressions. Although these methods achieve better decoding performance than the single-subject methods, they ignore learning inter-subject commonalities.



(a) Wills Aligner

(b) Three key components of Wills Aligner

Figure 2: Overview of the proposed Wills Aligner. Figure (a) illustrates the pipeline of our method. Figure (b) shows its three key components: Anatomical Alignment, Mixture of Brain Experts, and Learning Strategy.

Method

Formulation and Overview

Our target is to decode the visual information contained in the fMRI from a total of S subjects. Each fMRI dataset can be denoted as $D_s = \{(X_{s,i}, Y_{s,i}, L_{s,i}, I_{s,i})\}_{i=1}^{N_s}, s \in \{1, ..., S\}$, where $X_{s,i} \in \mathbb{R}^{L_s \times H_s \times W_s}$ represents fMRI, $Y_{s,i}$ is the seen image, $L_{s,i}$ is the image label, and $I_{s,i}$ is subject identity in the form of one-hot encoding. We explore visual decoding at three distinct levels: recognizing visual categories from fMRI data, performing bidirectional retrieval between fMRI and images, and reconstructing the seen image based on the corresponding fMRI.

As illustrated in Figure 2, our Wills Aligner comprises three key components: 1) **Anatomical Alignment**, an fMRI preprocessing to handle fMRI structural differences. 2) **Mixture of Brain Experts**, a subject-guided sparse mixture-of-expert (MoE) network for learning various fMRI patterns. 3) **Learning Strategy**, a two-phase training strategy to achieve multi-subject effective collaboration.

Anatomical Alignment

The brain structural differences among subjects lead to inconsistency in voxel counts and correspondence. Addressing this issue is a prerequisite for training with fMRI data from multiple subjects. To uniformly embed fMRI, we introduce the prior anatomical knowledge to assist fMRI alignment, i.e. anatomical alignment. The anatomical alignment process begins with the construction of the *fsaverage* standard brain template, a triangular surface mesh generated through spherical registration of 40 individual brain structures, based on the gray and white matter distributions, using an energy minimization algorithm. Then, a well-established mapping function maps any fMRI data to this template, preserving the cortical topological structure as much as possible, formally defined as $f : \mathbb{R}^{L_s \times H_s \times W_s} \rightarrow \mathbb{R}^{H_0 \times W_0}$, where $H_0 \times W_0$ represents the size of the standard brain template *fsaverage7* (Fischl et al. 1999). Due to the complete retention of brain activity and structure, this alignment strategy can achieve visual information losslessness. Subsequently, the one-dimensional sequence $\tilde{X}_{s,i} \in \mathbb{R}^d$ is obtained through the implementation of anatomical alignment, the selecting a region-of-interest (ROI), and the application of spatial flattening. After preprocessing, the aligned fMRIs are structure-standardized, but different fMRI patterns remain.

Mixture of Brain Experts

The fMRI pattern differences are the key factor hindering multi-subject collaboration. To allow one deep model to learn various fMRI patterns simultaneously, we introduce the Mixture of Brain Expert (MoBE) adapters. This branch network can recognize fMRI patterns and then learn using the corresponding brain experts. We parameterize the model $\mathcal{F}(\cdot)$ into three parts: θ is the subject-shared backbone, φ_s is MoBE adapters of Subject s, and r is the router.

MoBE adapters Similar to LoRA (Hu et al. 2021), MoBE adapters φ_s use a parallel structure connected to subject-shared parameters θ. Assuming that the model contains a total of M linear layers, thus there are $\theta = \{\theta^m\}_{m=1}^M$ and $\varphi_s = \{\varphi_s^m\}_{m=1}^M$. We use x^m to denote the input of the linear layer m and $x^1 = \tilde{X}_{s,i}$. Then, the output o^m of linear layer m can be formalized as:

$$o^m = (\theta^m + \sum_{s=1}^{S} \omega_s \varphi_s^m)\, x^m\,, \quad (1)$$

where ω_s represents the routing coefficient assigned to each MoBE adapter. To reduce the number of parameters in the MoBE adapters, we apply a low-rank decomposition to the parameters φ_s, similar to LoRA. In addition, we adopt the same initialization strategy as LoRA.

Subject-Guided Sparse Global Router Unlike existing MoE architectures that use separate routers for each MoE layer, we configure a subject-guided sparse global router. We find that the preprocessed fMRI data $\tilde{X}_{s,i}$ retains identifiable subject-specific fMRI patterns, allowing a simple model to classify. Therefore, we use subject identity $I_{s,i}$ as supervision, training an MLP to function as the router $\mathcal{R}(\cdot)$:

$$\mathcal{L}_{\text{router}} = \sum_{s=1}^{S} \sum_{i=1}^{N_s} \text{CrossEntropy}(\mathcal{R}(\tilde{X}_{s,i}; r), I_{s,i}) \ . \quad (2)$$

In our multi-subject visual decoding tasks, all MoBE adapters must act consistently and make the same routing choices when processing a specific subject's fMRI data. To avoid severe parameter redundancy, we configured a single router to handle routing for all MoBE adapters, i.e. a global router. To ensure the proper functioning of MoBE adapters, it is essential to implement a sparse routing strategy. we use the probability distributions output by the trained router as the routing coefficient:

$$[\omega_1, \omega_2, ..., \omega_S] = \mathcal{R}(\tilde{X}_{s,i}; r) \ . \quad (3)$$

Since the router/classifier can give extremely high classification confidence, this dense routing algorithm is equivalent to a sparse one. We will verify this in our experiments.

Learning Strategy

We split model training into two phases. In the first phase, we aim to guide the model backbone θ in learning the universal fMRI visual decoding knowledge. It is similar to a model pre-training and we achieve this commonality learning by a meticulously designed supervised objective. In the second phase, we aim to learn various fMRI patterns using MoBE adapters φ_s. We further employ a meta-learning strategy to integrate the learned fMRI patterns into semantics of deep representations. The supervised objective of downstream tasks is interspersed throughout both phases.

Downstream Task Learning We use $\mathcal{L}_{\text{task}}$ to represent downstream task supervision. It takes different forms depending on the visual decoding task. We calculate the fMRI representation $f_{s,i} = \mathcal{F}(\tilde{X}_{s,i}; \theta, \varphi_1, ..., \varphi_S, r)$.

For the classification task, we employ a classifier $\mathcal{G}(\cdot)$ and construct loss as follow:

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{cls}} = \sum_{s=1}^{S} \sum_{i=1}^{N_s} \text{CrossEntropy}(\mathcal{G}(f_{s,i}), L_{s,i}). \quad (4)$$

For the retrieval task, we first use the pre-trained ViT to extract the image representation $y_{s,i} = \text{ViT}(Y_{s,i})$. Then, we use a retrieval projector $\mathcal{H}(\cdot)$ to construct the bidirectional contrastive loss as follow:

$$\begin{aligned}
\mathcal{L}_{\text{fmri}} &= -\sum_{j=1}^{|B|} \log \frac{exp(\mathcal{H}(f_j)^\intercal \cdot y_j)}{\sum_{k=1}^{|B|} exp(\mathcal{H}(f_j)^\intercal \cdot y_k)} \ , \\
\mathcal{L}_{\text{image}} &= -\sum_{j=1}^{|B|} \log \frac{exp(y_j^\intercal \cdot \mathcal{H}(f_j))}{\sum_{k=1}^{|B|} exp(y_j^\intercal \cdot \mathcal{H}(f_k))} \ , \\
\mathcal{L}_{\text{task}} &= \mathcal{L}_{\text{retri}} = \mathcal{L}_{\text{fmri}} + \mathcal{L}_{\text{image}},
\end{aligned} \quad (5)$$

where B is a mini-batch sampled from entire fMRI.

For the image reconstruction task, we adopt the same basic pipeline as MindEye (Scotti et al. 2023). We map fMRI representations to the ViT image space through a diffusion prior $\mathcal{D}(\cdot)$. The reconstruction loss contains contrastive loss:

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{prior}}(\mathcal{D}(f_{s,i}), y_{s,i}) + \mathcal{L}_{\text{retri}} \ . \quad (6)$$

Commonality Learning Inter-subject commonality is universal knowledge used for fMRI visual decoding, which is supposed to be independent of subject identity. Essentially, it implies one-to-one correspondences between fMRI and images. Therefore, a natural insight is to increase the relevance of fMRI to the image and weaken its relevance to the subject's identity, which can be achieved through contrastive learning. To construct contrastive pairs, we utilize the similarity relation of representations. We represent fMRI representations, image representations, and subject identities in a mini-batch by $\mathbb{F}$, $\mathbb{Y}$, and $\mathbb{I}$, respectively. Then, we calculate their respective similarity relations by $\mathcal{M}_F = \mathbb{F}^\intercal\mathbb{F}$, $\mathcal{M}_Y = \mathbb{Y}^\intercal\mathbb{Y}$, and $\mathcal{M}_I = \mathbb{I}^\intercal\mathbb{I}$. Subsequently, We align the fMRI relation more closely with the image relation while distancing it from the subject identity relation, and term this process Semantic Relation Alignment (SRA):

$$\mathcal{L}_{\text{SRA}} = -\log \frac{\text{sim}(\mathcal{M}_F, \mathcal{M}_Y)}{\text{sim}(\mathcal{M}_F, \mathcal{M}_Y) + \text{sim}(\mathcal{M}_F, \mathcal{M}_I)} \ , \quad (7)$$

where $\text{sim}(\cdot)$ represents the cosine similarity. Based on this, our first training phase uses task loss and SRA loss to supervise the model backbone θ in learning commonality:

$$\theta^* = \arg\min_\theta \sum_{s=1}^{S} \sum_{i=1}^{N_s} (\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{SRA}}) \ . \quad (8)$$

Here α is the factor to balance two losses.

fMRI Pattern Learning We use a meta-learning strategy to learn various fMRI patterns and integrate them into semantics of deep representations. Formally, we define the fMRI data of each subject $\{D_s\}_{s=1}^{S}$ as support sets, and entire fMRI $D = \cup_{s=1}^{S} D_s$ as the query set. Then, the bi-level optimization of meta-learning can be formalized as:

$$\begin{aligned}
\theta^* &= \arg\min_\theta \sum_{\tilde{X} \in D} (\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{SRA}})(\tilde{X}; \theta, \varphi_1^*, ..., \varphi_S^*), \\
\text{s.t.} \quad \varphi_s^* &= \arg\min_{\varphi_s} \sum_{\tilde{X} \in D_s} \mathcal{L}_{\text{task}}(\tilde{X}; \varphi_s^*) \ .
\end{aligned} \quad (9)$$

The alternate method is more conducive to the model learning fMRI patterns. During each training session on support sets, MoBE adapters capture fMRI patterns. Subsequent

Method	# Model	# Parameters	mAP↑	AUC↑	Hamming↓
Single-Subject Vanilla Method	4	66M	0.258	0.854	0.033
Multi-Subject Vanilla Method	1	66M	0.150	0.767	0.039
EMB (Chehab et al. 2022)	1	66M	0.220	0.825	0.035
CLIP-MUSED (Zhou et al. 2024)	1	66M	0.258	0.877	0.030
Wills Aligners (**ours**)	1	**19M**	**0.424**	**0.937**	**0.024**

Table 1: Experimental results of the multi-label classification task. All baseline results are quoted from CLIP-MUSED (Zhou et al. 2024). We report the average value of four subjects, and our results are averaged over three runs.

Method	Objectives	# Parameters	Retrieval Accuracy	
			Image ↑	fMRI ↑
Mind Reader (Lin, Sprague, and Singh 2022)	InfoNCE Loss	2M	11.0%	49.0%
Brain Diffuser (Ozcelik and VanRullen 2023)	Contrastive Loss	3B	29.9%	21.4%
MindEye (Scotti et al. 2023)	Contrastive Loss	996M	83.7%	79.1%
MindEye (Scotti et al. 2023)	Contrastive Loss + MSE loss	996M	88.8%	84.9%
MindEye (Scotti et al. 2023)	SoftCLIP Loss + BiMixCo	996M	89.6%	82.2%
Wills Aligners (**ours**)	Contrastive Loss	**523M**	**95.4%**	**83.9%**

Table 2: Experimental results of the cross-modal retrieval on NSD. Our results are averaged over three runs. We report the results on Subj01, as previous works only provided results on Subj01.

fine-tuning on the query set integrates these fMRI patterns into the backbone network, which achieves improved deep representations and gives back fMRI pattern learning. Repeated execution of this alternating training will gradually improve multi-subject visual decoding performance.

Experiment

To validate the effectiveness and generalization of our Wills Aligner, we conduct extensive experiments, including classification, retrieval, reconstruction, and few-shot learning on the NSD dataset. Additionally, we perform further evaluations and ablation studies; please refer to our appendix[1].

Dataset

The Natural Scenes Dataset (NSD) is a massive 7T neuroscience dataset encompassing fMRI data. Throughout the NSD experiment, participants were presented with images sourced from MSCOCO (Lin et al. 2014), while their neural responses were recorded. Our study aligns with previous research by concentrating on Subj01, Subj02, Subj05, and Subj07, as these 4 subjects completed all 40 session scans.

Classification Experiment

The classification task stands as a coarse-grained brain visual decoding, which demands extracting and recognizing the categories of visual stimuli contained within fMRI data. We compare Wills Aligner with existing multi-subject classification methods on the NSD dataset. The implementation details follow CLIP-MUSED (Zhou et al. 2024). This task is a multi-label classification task, we employ three commonly used evaluation metrics in this field: mean Average Precision (mAP), the area under the receiver operating characteristic curve (AUC), and Hamming distance.

[1]https://arxiv.org/abs/2404.13282

Results and Analysis Table 1 shows the results of the classification experiment. The results indicate that our method achieves state-of-the-art performance in coarse-grained brain visual decoding, with a 64.3% improvement in mAP, a 6.8% improvement in AUC, and a 20.0% reduction in Hamming distance, while using fewer model parameters. The improvement can be attributed to two perspectives. On the one hand, our anatomical alignment is better than the baselines' fMRI alignment (such as PCA). Since their fMRI alignment focuses solely on the value itself without considering the brain structure, it tends to lose valuable information. In contrast, our fMRI alignment is based on the brain's anatomical knowledge, which retains useful brain information. On the other hand, our method achieves multi-subject collaboration so that visual decoding for the current subject benefits from other subjects' data. The similar performance between the single-subject vanilla method and CLIP-MUSED also suggests that existing methods merely avoid the damage to the performance caused by various fMRI patterns rather than exploiting the benefits of inter-subject commonality. Nevertheless, our method can achieve this.

Retrieval Experiment

The cross-modal retrieval between fMRI and images stands as a fine-grained brain visual decoding task, requiring fine-grained recognition of the visual semantics contained in the fMRI. Following MindEye (Scotti et al. 2023), we conduct bidirectional retrieval on NSD. We evaluate the performance using Top-1 retrieval accuracy. For image retrieval, we compute the cosine similarity between an fMRI representation and its respective ground truth image representation and 299 other randomly selected image representations in the test set. For each test sample, success is determined if the cosine similarity is greatest between the fMRI representation and its respective ground truth image representations (random chance

Few-Shot Ratio	Method	Subject 1		Subject 2		Subject 5		Subject 7	
		mAP ↑	AUC ↑	mAP ↑	AUC ↑	mAP ↑	AUC ↑	mAP ↑	AUC ↑
0.05 (1 session)	Vanilla	0.128	0.782	0.138	0.740	0.154	0.804	0.122	0.779
	Wills Aligner (**ours**)	**0.270**	**0.901**	**0.229**	**0.880**	**0.275**	**0.907**	**0.206**	**0.862**
0.1 (2 sessions)	Vanilla	0.143	0.808	0.168	0.834	0.171	0.827	0.138	0.802
	Wills Aligner (**ours**)	**0.322**	**0.915**	**0.287**	**0.898**	**0.334**	**0.921**	**0.276**	**0.888**
0.2 (4 sessions)	Vanilla	0.210	0.874	0.185	0.856	0.179	0.851	0.164	0.830
	Wills Aligner (**ours**)	**0.385**	**0.930**	**0.354**	**0.922**	**0.409**	**0.935**	**0.317**	**0.904**

Table 3: Experiment results of few-shot classification on NSD.

Few-Shot Ratio	Method	Subject 1		Subject 2		Subject 5		Subject 7	
		Image ↑	fMRI ↑	Image ↑	fMRI ↑	Image ↑	fMRI ↑	Image ↑	fMRI ↑
0.05 (1 session)	Vanilla	10.6%	1.9%	11.8%	1.2%	10.7%	1.8%	9.1%	1.2%
	Wills Aligner (**ours**)	**65.1%**	**47.9%**	**69.9%**	**49.3%**	**47.0%**	**29.3%**	**46.3%**	**30.0%**
0.1 (2 sessions)	Vanilla	26.5%	4.6%	29.5%	3.5%	20.0%	3.2%	20.9%	2.6%
	Wills Aligner (**ours**)	**75.9%**	**53.0%**	**77.4%**	**56.3%**	**53.8%**	**35.6%**	**54.8%**	**39.4%**
0.2 (4 sessions)	Vanilla	46.3%	13.8%	53.8%	12.8%	43.9%	8.5%	41.1%	8.6%
	Wills Aligner (**ours**)	**82.7%**	**61.1%**	**84.2%**	**64.2%**	**64.3%**	**46.6%**	**63.0%**	**46.8%**

Table 4: Experiment results of few-shot retrieval on NSD.

is 1/300). We repeat the evaluation for each test sample 30 times to account for the variability in the random sampling of batches. The same procedure is used for fMRI retrieval, except fMRI and images are flipped.

Results and Analysis Retrieval results are shown in Table 2. Our Wills Aligner outperforms MindEye, achieving a 15.4% performance improvement in image retrieval when the supervised objectives are equal. Even when MindEye uses BiMixCo data augmentation and SoftCLIP loss, it still outperforms MindEye, achieving a 7.8% improvement, while using fewer model parameters. Since all baselines are single-subject methods, there is no information loss of fMRI, and differential fMRI patterns are not a concern. Therefore, the results that more accurate visual information is decoded experimentally prove that preprocessed fMRI by anatomical alignment is almost equivalent to the original fMRI without losing visual information. In Addition, the performance improvement provided by Wills Aligner stems from multi-subject collaboration. Other subjects' data indeed helps fine-grained visual decoding for Subj01.

Few-Shot Visual Decoding Experiment

We perform the few-shot brain visual decoding experiment to further explore how much a single subject's visual decoding performance can benefit from multi-subject collaboration. The few-shot setting is more practical for real-world applications, as collecting large-scale fMRI data (40 sessions) for a single subject is challenging. Our experiments involve the classification and retrieval on NSD. We employ the few-shot setting for a given subject while the other subjects use the entire fMRI data for training. The few-shot ratios are set to be 0.05, 0.1, and 0.2, corresponding to 1, 2, and 4 sessions of fMRI. We compare our Wills Aligner with the single-subject vanilla method, where the model is trained exclusively on fMRI data from the few-shot subjects.

Results and Analysis Tables 3 and Table 4 show the performance of different few-shot subjects in the classification and retrieval tasks. The experiment results show that our Wills Aligner can significantly improve the brain visual decoding performance of few-shot subjects. In particular, for Subj01, we achieve a comparable image retrieval performance to MindEye (82.7% v.s. 83.7%) while we only use 4-session fMRI (1/10 of MindEye's training data). This suggests that our multi-subject collaboration strategy can transfer universal visual decoding knowledge from other subjects and improve the decoding performance of the current few-shot subject, highlighting the importance of collaboration.

Image Reconstruction Experiment

We further perform the fMRI-to-image reconstruction task on the NSD dataset to intuitively demonstrate the results of our visual decoding. Following previous studies (Scotti et al. 2023; Wang et al. 2024), we use eight image quality evaluation metrics. Among these, PixCorr, SSIM, AlexNet(2), and AlexNet(5) assess low-level perceptual aspects, while Inception, CLIP, EffNet-B, and SwAV evaluate high-level semantic aspects. We compare Wills Aligner with all existing methods. These baselines all used CLIP ViT-L/14 to extract image representations, leading to a fair comparison.

Quantitative Results and Analysis The quantitative experimental results are shown in Table 5. When compared with the state-of-the-art multi-subject reconstruction method, MindBridge, we achieve performance excellence in all metrics. Such an improvement stems from a better fMRI alignment strategy and multi-subject collaboration. Specifically, the anatomical alignment we use preserves useful fMRI information, while the max pooling employed by MindBridge may result in information loss. The fact that the single-subject MindBridge performs lower than MindEye also confirms the existence of information loss. In addition, our method can capture both inter-subject common-

Methods	# Models	Low-Level				High-Level			
		PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Mind-Diffuser (Lu et al. 2023)	4	0.254	**0.356**	94.2%	96.2%	87.2%	91.5%	0.775	0.423
MindEye (Scotti et al. 2023)	4	**0.309**	0.323	94.7%	97.8%	93.8%	94.1%	0.645	**0.367**
DREAM (Xia et al. 2024)	4	0.288	0.338	95.0%	97.5%	**94.8%**	**95.2%**	**0.638**	0.413
MindBridge (Wang et al. 2024)	4	0.148	0.259	86.9%	95.3%	92.2%	94.3%	0.713	0.413
MindBridge (Wang et al. 2024)	**1**	0.151	0.263	87.7%	95.5%	92.4%	94.7%	0.712	0.418
Wills Aligner (**ours**)	**1**	0.271	0.328	**95.8%**	**98.0%**	94.3%	94.8%	0.649	0.373

Table 5: Experiment results on fMRI-to-image reconstruction on NSD. Results are averaged over 4 subjects.

MoBEs	SRA	Classifiaction		Retrieval		Reconstruction							
		mAP ↑	AUC ↑	Image ↑	fMRI ↑	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
×	×	0.314	0.904	81.6%	77.9%	0.158	0.262	84.1%	93.6%	90.8%	92.6%	0.723	0.420
✓	×	0.417	0.935	92.3%	81.4%	**0.274**	0.327	94.5%	96.6%	91.3%	94.3%	0.689	0.382
×	✓	0.361	0.924	85.8%	79.1%	0.143	0.255	86.7%	94.2%	92.3%	94.4%	0.715	0.423
✓	✓	**0.424**	**0.937**	**95.4%**	**83.9%**	0.271	**0.328**	**95.8%**	**98.0%**	**94.3%**	**94.8%**	**0.649**	**0.373**

Table 6: Results of Ablation experiments. Results are averaged over 4 subjects.

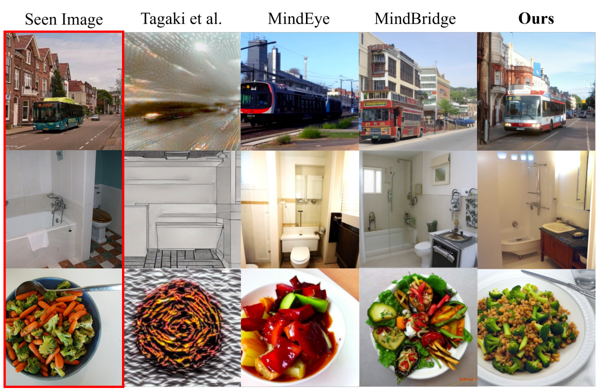


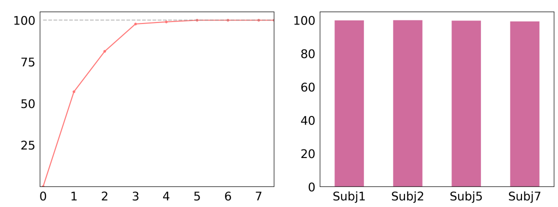
Figure 3: Reconstruction results for Subj01.

ality and different fMRI patterns, which MindBridge lacks. Even when compared to existing single-subject reconstruction methods, our Wills Aligner also achieves promising performance on some metrics. Since the multi-subject collaboration strategy tends to enhance the model's ability in visual semantic decoding, we fail to achieve a better performance than single-subject baselines on low-level metrics based on pixels, such as PixCorr and SSIM.

Qualitative Results Figure 3 visualizes the results of our image reconstruction. It can be seen that our reconstruction is more advantageous in terms of semantic accuracy compared to MindBridge. For instance, in the top row, our reconstruction of the "bus" exhibits a better spatial position, which is more similar to the seen image. Similarly, in the bottom row, we have accurately reconstructed "broccoli" and "cutlery". These observations further explain why our quantitative metrics are better than baselines.

Exploratory Experiments

Ablation Experiments We perform ablation experiments to explore the performance sources of our Wills Aligner. We ablate two key components of our Wills Aligner. One is MoBE adapters together with the corresponding meta-learning strategy, and the other is the SRA loss used for learning inter-subject commonality. The experimental results are shown in Table 6, indicating that both key components of Wills Aligner are useful for performance improvement. We also find that SRA loss does not improve performance as much as MoBE adapters, suggesting fMRI pattern learning is indispensable in multi-subject collaboration.



(a) Routing Accuracy (b) Routing Confidence

Figure 4: Analysis on router. The left figure shows the growth of the router's test set classification accuracy with training steps. The right figure shows the average probabilities of the router's top-1 prediction on the test set.

Analysis on Router Figure 4 illustrates the analysis of the router. It is evident that, despite anatomical alignment, fMRI from different subjects exhibit distinct fMRI patterns, allowing recognition by a simple MLP. After training, the router achieves 100% accuracy with high confidence, demonstrating that our router achieves sparse routing in fact.

Conclusion

In this paper, we propose Wills Aligner. It achieves multi-subject collaborative brain visual decoding by anatomical alignment, learning inter-subject commonality, and learning various fMRI patterns. We have substantiated its effectiveness and generalization through extensive experiments. Future works could consider fusing fMRI data from different subjects to achieve unbiased visual decoding or use trained decoding models to assist neuroscience research.

Acknowledgements

This research is supported by the National Key Research and Development Program of China (No. 2022YFB3104700), the National Natural Science Foundation of China (No. 62376198, No. 62076182), Shanghai Baiyulan Pujiang Project (No. 08002360429).

References

Allen, E. J.; St-Yves, G.; Wu, Y.; Breedlove, J. L.; Prince, J. S.; Dowdle, L. T.; Nau, M.; Caron, B.; Pestilli, F.; Charest, I.; et al. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1): 116–126.

Chehab, O.; Défossez, A.; Jean-Christophe, L.; Gramfort, A.; and King, J.-R. 2022. Deep Recurrent Encoder: an end-to-end network to model magnetoencephalography at scale. *Neurons, Behavior, Data analysis, and Theory*.

Chen, Z.; Qing, J.; Xiang, T.; Yue, W. L.; and Zhou, J. H. 2023. Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22710–22720. IEEE.

Chen, Z.; Qing, J.; and Zhou, J. H. 2023. Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity. In *Advances in Neural Information Processing Systems*.

Destrieux, C.; Fischl, B.; Dale, A.; and Halgren, E. 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, 53(1): 1–15.

Du, C.; Fu, K.; Li, J.; and He, H. 2023. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ferrante, M.; Boccato, T.; and Toschi, N. 2023. Through their eyes: multi-subject Brain Decoding with simple alignment techniques. *arXiv preprint arXiv:2309.00627*.

Fischl, B.; Sereno, M. I.; Tootell, R. B.; and Dale, A. M. 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, 8(4): 272–284.

Gong, Z.; Bao, G.; Zhang, Q.; Wan, Z.; Miao, D.; Wang, S.; Zhu, L.; Wang, C.; Xu, R.; Hu, L.; Ke, L.; and Yu, Z. 2024a. NeuroClips: Towards High-fidelity and Smooth fMRI-to-Video Reconstruction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Gong, Z.; Zhang, Q.; Bao, G.; Zhu, L.; Liu, K.; Hu, L.; and Miao, D. 2024b. MindTuner: Cross-Subject Visual Decoding with Visual Fingerprint and Semantic Correction. *arXiv preprint arXiv:2404.12630*.

Gong, Z.; Zhang, Q.; Bao, G.; Zhu, L.; Zhang, Y.; LIU, K.; Hu, L.; and Miao, D. 2024c. Lite-Mind: Towards Efficient and Robust Brain Representation Learning. In *ACM Multimedia 2024*.

Horikawa, T.; and Kamitani, Y. 2017. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1): 15037.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models.

Huo, J.; Wang, Y.; Wang, Y.; Qian, X.; Li, C.; Fu, Y.; and Feng, J. 2025. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, 56–73. Springer.

Kamitani, Y.; and Tong, F. 2005. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5): 679–685.

Li, C.; Qian, X.; Wang, Y.; Huo, J.; Xue, X.; Fu, Y.; and Feng, J. 2025. Enhancing Cross-Subject fMRI-to-Video Decoding with Global-Local Functional Alignment. In *European Conference on Computer Vision*, 353–369. Springer.

Lin, S.; Sprague, T.; and Singh, A. K. 2022. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35: 29624–29636.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, volume 8693, 740–755.

Linden, D. 2021. Section 3 - Introduction. In Hampson, M., ed., *fMRI Neurofeedback*, 161–169. Academic Press. ISBN 978-0-12-822421-2.

Liu, Y.; Ma, Y.; Zhou, W.; Zhu, G.; and Zheng, N. 2023. BrainCLIP: Bridging Brain and Visual-Linguistic Representation via CLIP for Generic Natural Visual Stimulus Decoding from fMRI. *arXiv preprint arXiv:2302.12971*.

Lu, Y.; Du, C.; Zhou, Q.; Wang, D.; and He, H. 2023. MindDiffuser: Controlled Image Reconstruction from Human Brain Activity with Semantic and Structural Diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5899–5908.

Naselaris, T.; Kay, K. N.; Nishimoto, S.; and Gallant, J. L. 2011. Encoding and decoding in fMRI. *Neuroimage*, 56(2): 400–410.

Nie, X.; Hu, B.; Gao, X.; Li, L.; Zhang, X.; and Xiao, B. 2023. BMI-Net: A Brain-inspired Multimodal Interaction Network for Image Aesthetic Assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5514–5522.

Ozcelik, F.; and VanRullen, R. 2023. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. *arXiv preprint arXiv:2303.05334*.

Palazzo, S.; Spampinato, C.; Kavasidis, I.; Giordano, D.; Schmidt, J.; and Shah, M. 2020. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11): 3833–3849.

Parthasarathy, N.; Batty, E.; Falcon, W.; Rutten, T.; Rajpal, M.; Chichilnisky, E. J.; and Paninski, L. 2017. Neural Networks for Efficient Bayesian Decoding of Natural Images from Retinal Neurons. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, 6434–6445.

Pereira, F.; Lou, B.; Pritchett, B.; Ritter, S.; Gershman, S. J.; Kanwisher, N.; Botvinick, M.; and Fedorenko, E. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1): 963.

Quan, R.; Wang, W.; Tian, Z.; Ma, F.; and Yang, Y. 2024. Psychometry: An omnifit model for image reconstruction from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 233–243.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Roelfsema, P. R.; Denys, D.; and Klink, P. C. 2018. Mind reading and writing: the future of neurotechnology. *Trends in cognitive sciences*, 22(7): 598–610.

Scotti, P. S.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Cohen, E.; Dempster, A. J.; Verlinde, N.; Yundler, E.; Weisberg, D.; et al. 2023. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. *arXiv preprint arXiv:2305.18274*.

Scotti, P. S.; Tripathy, M.; Villanueva, C. K. T.; Kneeland, R.; Chen, T.; Narang, A.; Santhirasegaran, C.; Xu, J.; Naselaris, T.; Norman, K. A.; et al. 2024. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. *arXiv preprint arXiv:2403.11207*.

Shen, G.; Horikawa, T.; Majima, K.; and Kamitani, Y. 2019. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1): e1006633.

Takagi, Y.; and Nishimoto, S. 2023. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14453–14463.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, S.; Liu, S.; Tan, Z.; and Wang, X. 2024. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11333–11342.

Xia, W.; de Charette, R.; Oztireli, C.; and Xue, J.-H. 2024. Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 8226–8235.

Zhang, Y.; Bu, T.; Zhang, J.; Tang, S.; Yu, Z.; Liu, J. K.; and Huang, T. 2022. Decoding pixel-level image features from two-photon calcium signals of macaque visual cortex. *Neural Computation*, 34(6): 1369–1397.

Zhou, Q.; Du, C.; Wang, S.; and He, H. 2024. CLIP-MUSED: CLIP-Guided Multi-Subject Visual Neural Information Semantic Decoding.