# Lite-Mind:
# Towards Efficient and Robust Brain Representation Learning

Anonymous Authors

## ABSTRACT

The limited data availability and the low signal-to-noise ratio of fMRI signals lead to the challenging task of fMRI-to-image retrieval. State-of-the-art MindEye remarkably improves fMRI-to-image retrieval performance by leveraging a large model, i.e., a 996M MLP Backbone per subject, to align fMRI embeddings to the final hidden layer of CLIP's Vision Transformer (ViT). However, significant individual variations exist among subjects, even under identical experimental setups, mandating the training of large subject-specific models. The substantial parameters pose significant challenges in deploying fMRI decoding on practical devices. To this end, we propose **Lite-Mind**, a lightweight, efficient, and robust brain representation learning paradigm based on Discrete Fourier Transform (DFT), which efficiently aligns fMRI voxels to fine-grained information of CLIP. We elaborately design a DFT backbone with Spectrum Compression and Frequency Projector modules to learn informative and robust voxel embeddings. Our experiments demonstrate that **Lite-Mind** achieves an impressive 94.6% fMRI-to-image retrieval accuracy on the NSD dataset for Subject 1, with 98.7% fewer parameters than MindEye. **Lite-Mind** is also proven to be able to be migrated to smaller fMRI datasets and establishes a new state-of-the-art for zero-shot classification on the GOD dataset.

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → *Computational biology*; • **Human-centered computing** → *Human computer interaction (HCI)*.

## KEYWORDS

fMRI, Brain-computer Interface (BCI), Cross-modal Retrieval, Efficient Model

## 1 INTRODUCTION

Brain decoding holds immense significance in elucidating intricate mental processes and serves as a pivotal technology for brain-computer interfaces [14, 25, 27]. Among diverse brain decoding tasks, the decoding of visual information stands out as a paramount yet challenging endeavor, allowing us to unravel the complex mechanisms involved in visual processing, object recognition, scene understanding, and even higher-order cognitive functions. In the pursuit of decoding natural visual information, functional magnetic
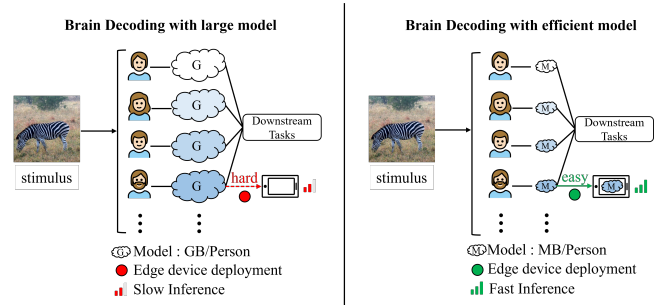
**Figure 1: The significance of lightweight brain decoding model. Subject-specific models are visually represented by clouds of distinct colors, and the cloud size correlates with model parameters.**

resonance imaging (fMRI) is widely employed as a non-invasive modality to decipher the perceptual and semantic information intricately encoded within the cerebral cortex [21]. It draws considerable attention to fMRI-to-image retrieval/reconstruction tasks. This paper focuses primarily on tasks related to fMRI-to-image retrieval.

The success of fMRI-to-image retrieval heavily relies on aligning fMRI signals with the representation space of images. However, fMRI data often suffers from spatial redundancy, noise, and sample sparsity, leading to poor representation of fMRI signals and potential overfitting to noise distribution for deep models. To address these challenges, previous studies have found ridge regression and shallow linear models to be effective solutions for fMRI-based models [1, 9, 13, 15, 28, 46]. Recently, it has been demonstrated the effectiveness of leveraging pretrained CLIP models [29] as a powerful bridge between fMRI voxels and images, since CLIP's image embeddings capture fine-grained and semantic information of pictures [20, 22, 24, 34]. Initially, the CLS embeddings of the CLIP image encoder was used to align voxels in Mind-Reader [20]. Considering the CLS embeddings alone provide limited information for fMRI data, BrainClip [22] and UniBrain [24] introduce the text annotations of images to assist in aligning fMRI voxels with image representations. Brain-diffuser [26] aligns voxel representations with the last hidden layer of CLIP using large-scale linear models ($257 \times 14M$) for the first time.

Among the CLIP-related models, MindEye [34] stands out by employing a large-scale (996M) MLP Backbone and Projector to map fMRI voxels and align their representations with the final hidden layer of CLIP using contrastive learning. MindEye achieves remarkable breakthroughs, surpassing 90% accuracy for both image and brain retrieval, outperforming previous state-of-the-art accuracies that were below 50%. However, the practical application of large-scale models in brain science research faces limitations. Each subject's fMRI data can exhibit significant variations, even within identical experimental setups, making it currently infeasible to establish a universal encoding model that performs well for

all individuals. Consequently, training individual encoding models for each subject becomes necessary, as depicted in Figure 1. Apparently, it is impractical to expect hospitals to have access to computing resources with sufficient performance to train a unique large-scale model (e.g., MindEye) for each individual. Therefore, the ideal scenario involves equipping individuals with lightweight brain-computer interface edge devices, enabling image representations to be efficiently retrieved by a lightweight model. To this end, there is an urgent need to develop a brain network that is lightweight and efficient, as illustrated in the right segment of Figure 1, for practical implementation and deployment of brain-computer interfaces in real-world scenarios.

MindEye strongly benefits from an MLP backbone with a high parameter count, which allows for straightforward and powerful compression and preservation. However, when dealing with fMRI data, which typically has a low signal-to-noise ratio, an MLP focusing on voxel value-wise mappings may inadvertently suppress informative voxel signal values while being inefficient at reducing noisy values. This can lead to model parameter redundancy and decelerated convergence speed. Recently, the Fourier Transform has gained attention in deep learning, demonstrating lightweight characteristics and improved efficiency in learning frequency domain patterns, due to its global perspective and energy concentration properties [45]. Additionally, the Fourier Transform naturally possesses the advantage of processing signals: it is more effective and efficient to filter out noise and maintain informative voxel signals in the frequency domain. It presents a promising alternative to large-scale MLPs for encoding fMRI signals.

In order to mitigate the noise effect of fMRI and obtain more robust latent representations, we designed global filter-based Filter Blocks for fMRI denoising and fMRI spectrum compression. In addition to the global view property of frequency domain filtering, which can block out the noise and thus obtain better robust representations, the frequency domain computation also brings a great efficiency improvement compared to the large MLP structure. Since the features of the compressed fMRI we obtain are more concentrated in the frequency domain, we also design a Frequency Projector based on MLP in the frequency domain. Combining all our computational paradigms in frequency domain, we propose **Lite-Mind**, a lightweight, efficient, and robust brain representation learning paradigm. We redesign MindEye with our elaborate Discrete Fourier Transform (DFT) Backbone, avoiding the large MLP Backbone used in MindEye. Extensive experiments show that **Lite-Mind** achieves 94.6% retrieval accuracy for Subject 1 on the NSD dataset, with 98.7% fewer parameters than MindEye. **Lite-Mind** also proves its adaptability to smaller brain datasets and establishes a new state-of-the-art zero-shot classification on the GOD dataset. Our contributions are summarized as follows:

- We demonstrated that the Fourier Transform has the advantage of being lightweight and efficient for fMRI representation learning in the frequency domain.
- We proposed a novel **DFT Backbone** with elaborate Spectrum Compression and Frequency Projector modules, which is theoretically and empirically verified lightweight, efficient, and robust for brain representation.

- We migrated **Lite-Mind** to various downstream tasks, and demonstrated its robustness. It achieves state-of-the-art zero-shot classification performance on the GOD dataset.

## 2 RELATED WORK

### 2.1 Brain Visual Decoding

The study of visual decoding using fMRI in the human brain has been a long-standing endeavor. In the most important fMRI-image reconstruction and partial fMRI-image retrieval tasks, the foundation is how to align fMRI signals to the image representation space or intermediate space. Due to the low signal-to-noise ratio of fMRI, the initial research emphasized using models based on linear regression to extract fMRI signals and images into the intermediate space for functional decoding of the human brain or reconstruction of faces and natural scenes [37]. Further research utilizes pre-trained VGG to enrich hierarchical image features [7, 13, 35]. With the development of cross-modal tasks, the image-text pre-training model CLIP has been introduced into fMRI-image research. Mind-reader [20] is the first to adopt a contrastive learning approach, using the *nsdgenal* ROI region of NSD data and a shallow Resnet-like model to align fMRI with the CLS embeddings output by the CLIP image encoder. Mind-Reader did not achieve good performance in fMRI-image retrieval when using InfoNCE loss. BrainClip [22] uses a VAE model for retrieval tasks while still using the CLS embeddings. Brain diffuser [26] introduced a fine-grained representation of the image of the last hidden layer in CLIP. However they used 257 different linear regression models to align each layer of the 257 hidden layers, resulting in the model being too cumbersome. Mind-Vis [5] conducted self-supervised mask pre-training on fMRI signals for another dataset, BOLD5000, and demonstrated that voxel patches can be used to process fMRI voxel values. MindEye [34] uses a large MLP backbone for contrastive learning, aligning the last hidden layer $257 \times 768$ of CLIP, and using diffusion prior in DALLE·2 [30] for the first time to narrow the disjointed vectors. At the same time, it creatively proposed a mixed contrast loss MixCo as a data augmentation method to make the large model converge. The recent trend in brain fMRI decoding research tends to use larger models to achieve better performance on downstream tasks [16, 41], ignoring the privacy and deployment issues discussed above. However, to the best of our knowledge, no previous research has focused on lightweight and efficient brain networks.

### 2.2 Fourier Transform in Deep Learning

Fourier Transform plays a vital role in the area of digital signal processing. It has been introduced to deep learning for enhanced learning performance [8, 11, 44], and has demonstrated excellent performance in many fields, such as computer vision, natural language processing, and time series analysis. In the field of computer vision, GFNet [31] utilizes Fast Fourier Transform(FFT) to convert images to the frequency domain and exchange global information between learnable filters. As a continuous global convolution independent of input resolution, Guibas et al. [12] design the Adaptive Fourier Neural Operator(AFNO) frame token mixing. Xu et al. [42] devise a learning-based frequency selection method to identify trivial frequency components and improve image classification accuracy. As for natural language processing, Lee-Thorp et al. [19] use
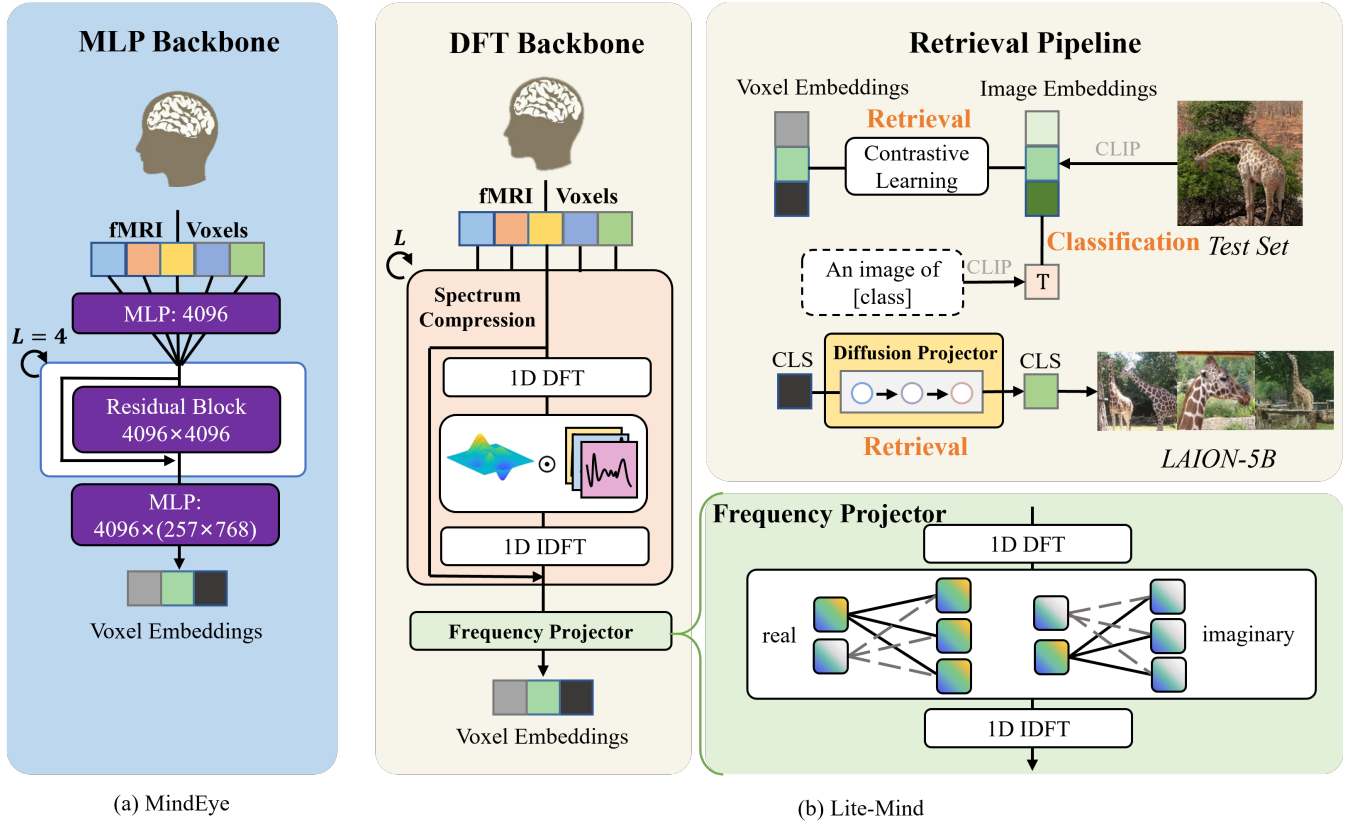
(a) MindEye

(b) Lite-Mind

**Figure 2: Overview of our Lite-Mind. Figures (a) and (b) show the architecture of the MLP Backbone of MindEye and the DFT backbone and Retrieval Pipeline, respectively. fMRI voxels are inputted into DFT Backbone to obtain voxel embeddings.**

the Fourier Transform as a text token mixing mechanism for text classification. In research of time series forecast, Fourier Transform has deeper applications [3, 17, 18, 43]. To increase the accuracy of multivariate time-series forecasting, Cao et al. [3] propose a spectral temporal graph neural network (StemGNN), which mines the correlations and time dependencies between sequences in the frequency domain. Yang et al. [43] propose bilinear temporal spectral fusion (BTSF), which updates the feature representation in a fused manner by explicitly encoding time-frequency pairs and using two aggregation modules: spectrum-to-time and time-to-spectrum. Yi et al. [45] prove that frequency-domain MLP is more efficient than time-domain MLP. Processing fMRI frequency domain information is more helpful for understanding the mechanism of collaborative operation within the human brain. However, there is currently no research on processing fMRI signals in the frequency domain for representation learning.

## 3 LITE-MIND

### 3.1 Overview

As illustrated in Figure 2, Lite-Mind comprises two main components: 1) **DFT Backbone** consists of *fMRI Spectrum Compression* and *Frequency Projector*, mapping flattened voxels to an intermediate space. 2) **Retrieval pipeline** contains *Diffusion Projector*, *Contrastive Learning*, and *Downstream tasks* with the voxel embeddings from the DFT backbone.

### 3.2 DFT Backbone

Sampling an fMRI-image pair $(x, y)$ from dataset $D$, $x$ denotes the fMRI data and $y$ denotes the paired image.

*3.2.1 fMRI Spectrum Compression.* **Patchify and Embedding** The fMRI data $x$ after ROI preprocessing and flattening is a one-dimensional long vector composed of voxels. We conduct patchify firstly because it has been verified as simple and effective for spatial/temporal representation of long sequences or high-dimensional vectors. At first, we divide $x$ into $n$ non-overlapping patches $x = [x_1, x_2, ..., x_n]$ using zero-padding, and adopt convolutions with positional encoding to obtain multiple tokens $t = [t_1, t_2, ..., t_n]$.

**Filter Block.** The spectrum of voxel tokens is obtained by 1D DFT as follows:

$$X[k] = F(t) = \sum_{i=1}^{n} t_i e^{-ki(2\pi/n)j} \tag{1}$$

Where $X \in \mathbb{C}^{n \times d}$ is a complex tensor, $X[k]$ is the spectrum of $t$ at the frequency $2\pi k/n$, $F(\cdot)$ denotes the 1D DFT along the voxel token dimension. $i$ denotes the $i$-th token, and $j$ denotes unit imaginary.

Voxel spatial features are effectively consolidated within each element of the frequency spectrum of voxel tokens, enabling the extraction of informative features from voxels through the point-wise product in the frequency domain. Accordingly, we introduce learnable filters $\mathbf{K} = [\mathbf{k}_1, \mathbf{k}_2, ..., \mathbf{k}_M]$ to filter and compress the spectrum as follows, where $\mathbf{K}$ denotes the filter library and $M$ is

the number of filters in the library.

$$\hat{X} = \sum_{m=1}^{M} \frac{1}{n} |X|^2 \odot \mathbf{k}_m cos(\frac{(2m-1)\pi}{2M})$$ (2)

where $\odot$ is the element-wise multiplication, $|X|^2$ is the power spectrum of $X$. The $|X|^2$ operation smooths the spectrum, highlighting the main components of the fMRI spectrum from a global perspective. $cos((2m-1)\pi/2M)$ adopts the filtering form of Discrete Cosine Transform (DCT), which compacts better energy and can aggregate the significant information in fMRI voxels. Its combination with the channel filter library **K** allows for efficient feature compression and noise denoising for fMRI from the frequency domain.

Finally, we employ Inverse Discrete Fourier Transform (IDFT) $F^{-1}(\cdot)$ to convert the spectrum back into the spatial domain:

$$\hat{t} \leftarrow F^{-1}(\hat{X})$$ (3)

*3.2.2 Frequency Projector.* The fMRI spectrum after filtering has a more concentrated feature in the frequency domain compared to the spatial domain, which is caused by the energy concentration effect of the frequency domain calculation. We therefore designed a Frequency Projector in the frequency domain to align the image tokens, which has the added benefit of maintaining all computations of our DFT Backbone in the frequency domain, where lightness and efficiency are guaranteed. To align voxel tokens to image tokens, we mapped the filtered and compressed fMRI tokens by an MLP-like projector in the frequency domain(FreMLP). Similarly, we follow Equation (2) to convert $\hat{t}$ to the frequency domain.

**FreMLP.** Following Equation (2), the complex number output of $\hat{t}$ after DFT is $\hat{X} \in \mathbb{C}^{n \times d}$. Given a complex number weight matrix $\mathcal{W} \in \mathbb{C}^{n \times n'}$, and a complex number bias $\mathcal{B} \in \mathbb{C}^{n'}$, then the FreMLP can be formulated as:

$$X' = \sigma(\hat{X}^T \mathcal{W} + \mathcal{B})^T$$ (4)

Where $X' \in \mathbb{C}^{n' \times d}$ is the final output, and $\sigma$ is the activation function. As both $\hat{X}$ and $\mathcal{W}$ are complex numbers, according to the rule of multiplication of complex numbers(see Appendix D.1 for details), we further extend the Equation (5) to:

$$X' = (\sigma(Re(\hat{X}^T)\mathcal{W}_r - Im(\hat{X}^T)\mathcal{W}_i + \mathcal{B}_r) \\ + j\sigma(Re(\hat{X}^T)\mathcal{W}_i + Im(\hat{X}^T)\mathcal{W}_r + \mathcal{B}_i))^T$$ (5)

Where $Re(\cdot)$ denotes the real part of $\hat{X}^T$, $Im(\cdot)$ denotes the imaginary part, $\mathcal{W} = \mathcal{W}_r + j\mathcal{W}_i$ and $\mathcal{B} = \mathcal{B}_r + j\mathcal{B}_i$. Due to the more significant features of filtered and compressed tokens in the frequency domain, we replace the final FC layer of MindEye's MLP Backbone with FreMLP. For more theoretical proof and implementation of FreMLP, please refer to Appendix D.1. Finally, we follow IDFT $t' \leftarrow F^{-1}(X')$ to get final tokens $t'$, and use voxel embedding $f$ to represent them.

## 3.3 Retrieval Pipeline

*3.3.1 Optimization Objective.* The optimization objective for the retrieval task is as follows:

$$\omega^* = \underset{\omega}{argmax} \sum_{(x,y) \in D} SIM(DFT(x; \omega), CLIP(y))$$ (6)

where $\omega$ is the weight of our DFT backbone, $SIM(\cdot)$ denotes the cosine similarity, and $CLIP(\cdot)$ denotes using CLIP model to extract image embedding.

*3.3.2 Diffusion Projector.* When searching on a large-scale dataset like LAION-5B, image-to-image performs better than fMRI-to-image and can find more similar images. Therefore, we used a diffusion projector to translate the disjointed voxel CLS embeddings of the DFT Backbone output, using a DALLE · 2[30] similar diffusion model to obtain the image CLS embeddings and retrieve it on LAION-5B: $\mathcal{V}' = Diffusion(f)$. The Diffusion Projector consists of a transformer for image CLS embedding forecasting.

*3.3.3 Contrastive Learning.* Due to the inherent differences between different modalities, we use CLIP's contrastive loss to train the DFT backbone. Sampling a batch $B$ from voxel-image pairs, the contrastive loss is defined below:

$$L_{contr} = -\frac{1}{|B|} \sum_{s=1}^{|B|} \log \frac{\exp(f_s^\top \cdot \mathcal{V}_s/\tau)}{\sum_{i=1}^{|B|} \exp(f_s^\top \cdot \mathcal{V}_i/\tau)},$$ (7)

where $f_s$ is the $s$-th voxel embeddings, $\mathcal{V}_s$ is its corresponding image embeddings, $\tau$ is a temperature factor.

To perform image-to-image retrieval on LAION-5B, we use MSE loss to constrain the generation of approximate image embeddings:

$$L_{mse} = \frac{1}{|B|} \sum_{s=1}^{|B|} ||\mathcal{V}_s - \mathcal{V}'_s||_2^2$$ (8)

Accordingly, the training loss is defined below:

$$L = L_{contr} + \alpha L_{mse}$$ (9)

In general, when $\alpha = 0$, Lite-Mind is trained to output voxel embeddings for retrieval on the NSD test set. However, for LAION-5B retrieval, $\alpha$ is non-zero to map the voxel CLS embeddings outputted by Lite-Mind to image CLS embeddings, enabling online retrieval.

*3.3.4 Downstream tasks.* The retrieval process on various downstream tasks is shown in the upper right of Figure 2.

**Test set retrieval.** The cosine similarities between voxel embeddings from DFT Backbone and image embeddings from CLIP ViT are directly calculated.

**LAION-5B retrieval.** Voxel CLS embeddings from the DFT backbone are translated to image CLS embeddings through the Diffusion Projector for online LAION-5B retrieval.

**Zero-shot classification.** Image retrieval is performed on novel classes on the test set, and the similarities between the retrieved images and simple CLIP class text prompts are calculated for classification tasks.

## 4 EXPERIMENTS

### 4.1 Dataset

Natural Scenes Dataset (NSD) is an extensive 7T fMRI dataset gathered from 8 subjects viewing images from MSCOCO-2017 dataset [1], which contains images of complex natural scenes. Participants viewed three repetitions of 10,000 images with a 7-Tesla fMRI scanner over 30–40 sessions. More details can be found on the NSD official website[1]. Our experiments focused on Subj01, Subj02,

---
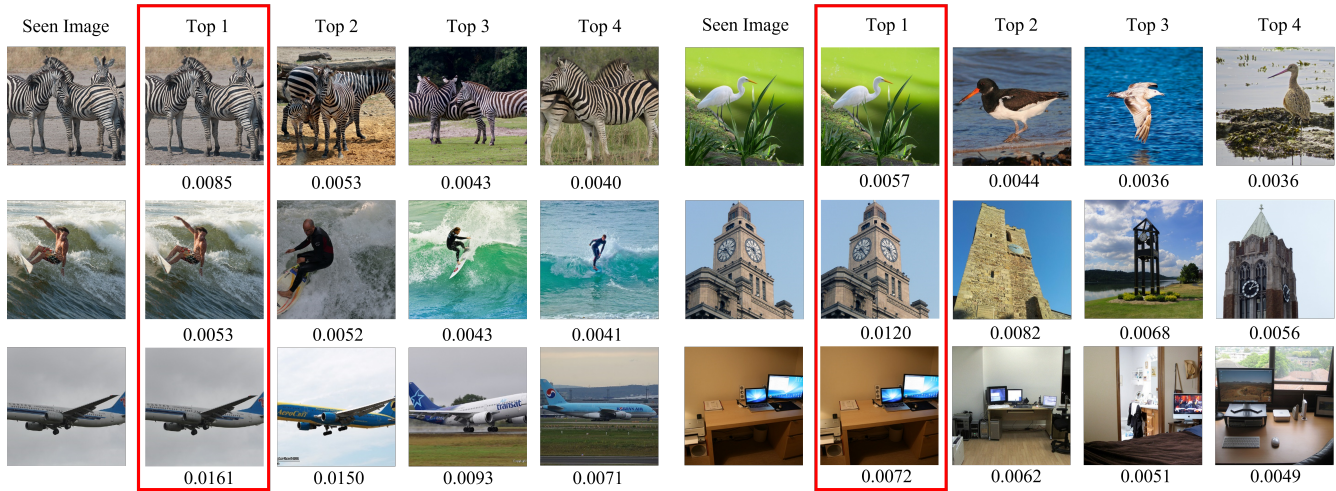
[1] https://naturalscenesdataset.org

**Figure 3: Partial retrieval results of Lite-Mind on all 982 test images for Subject 1. With 12.5M DFT Backbone, Lite-Mind can still find the exact Top-1 image pair from the test set of 982 images with 89.3% accuracy and 95.5% accuracy of image-fMRI retrieval(random chance = 0.1%) and can distinguish among confusable candidates. The number below each image represents the similarity score. See more cases including failure retrieval in Appendix C.4.**

Subj05, and Subj07, who finished all viewing trials. Each subject's training set comprises 8859 image stimuli and 24980 fMRI trials (with the possibility of 1-3 repetitions per image). The test set contains 982 image stimuli and 2770 fMRI trials. Responses for images with multiple fMRI trials are averaged across these trials. By applying the *nsdgeneral* ROI mask with a 1.8 mm resolution, we obtained ROIs for the four subjects, comprising 15724, 14278, 13039, and 12682 voxels respectively. These regions span visual areas ranging from the early visual cortex to higher visual areas. Our experimental setup is consistent with the NSD image reconstruction and retrieval articles [20, 24, 26, 34, 37].

Generic Object Decoding (GOD) Dataset[2] was created by Horikawa and Kamitani [13], consisting of fMRI recordings of five healthy subjects who were presented with images from ImageNet [6]. The GOD Dataset includes 1250 distinct images selected from 200 ImageNet categories. Among these, 1200 training images are drawn from 150 categories, and 50 test images are from the remaining 50 categories. The training and test image stimuli were presented to the subjects once and 35 times respectively, resulting in 1200 and 1750 fMRI instances. We used preprocessed ROIs, encompassing voxels from the early visual cortex to higher visual areas. For each test image, the fMRI responses from different trials were averaged.

### 4.2 Implementation details

To fairly compare different backbones, we did not use any data augmentation methods, such as voxel mixing loss MixCo and image slicing enhancement. Due to MindEye only disclosing the MLP Backbone performance of Subject 1, only the performance of Subject 1 is provided in Table 1a. The experimental results of more subjects are shown in Appendix B.1. All of our experiments were conducted on a single Tesla V100 32GB GPU. More experimental details and hyperparameter settings can be found in Appendix A.

---

[2]http://brainliner.jp/data/brainliner/Generic_Object_Decoding

## 5 RESULTS

### 5.1 fMRI/image retrieval

Image retrieval refers to retrieving the image embeddings with the highest cosine similarity based on voxel embeddings on the test set. If a paired image embedding is retrieved, the retrieval is considered correct. fMRI retrieval is the opposite process mentioned above. Note that there are many semantically and visually similar images on the NSD test set, considered to be similar in CLIP space. Whether the model can correctly retrieve as MindEye tests the fine-grained alignment ability to the image. For test retrieval, we adhered to the identical methodology as Lin et al. [20] and MindEye [34] to compute the retrieval metrics presented in Table 1a. For each test sample, we randomly selected 299 images from the remaining 981 images in the test set and calculated the cosine similarity between the voxel embeddings and 300 images. The retrieval accuracy refers to the proportion of successful retrieval of corresponding images in the 982 voxel embeddings of the test set. We adjusted the random number seed of 30 randomly selected images to average the accuracy of all samples. The experimental results are shown in Table 1a. We also conducted retrieval experiments on the remaining three subjects to demonstrate the universality of DFT Backbone. The detailed results and discussions are in Appendix B.2.

As shown in Table 1a, compared to MLP Backbone, our DFT Backbone improves retrieval accuracy by 5% and 14.9% for two retrieval ways, indicating that the fine-grained representation of the image comes from the rich representation of the last hidden layer of CLIP rather than MLP's excessive attention to each voxel value. Note that our DFT Backbone is closest to MindEye's MLP Backbone with only contrastive learning, while Prior in MLP Backbone+Prior includes MSE loss for image reconstruction.

| Method | Model | Parameters | Retrieval | |
|---|---|---|---|---|
| | | | Image↑ | Brain↑ |
| Lin et al. [20] | deep models | 2.34M | 11.0% | 49.0% |
| Ozcelik... [26] | 257 separate linear regression models | 3B | 29.9% | 21.4% |
| MindEye [34] | MLP Backbone | 940M | 89.6% | 82.2% |
| MindEye [34] | MLP Backbone+Projector | 996M | 88.8% | 84.9% |
| MindEye [34] | MLP Backbone+Prior | 1.2B | 93.4% | 90.1% |
| Lite-Mind(ours) | DFT Backbone | **12.5M** | **94.6%** | **97.1%** |

(a) Quantitative comparison of Lite-Mind retrieval performance against other models for Subject 1. Image retrieval refers to the hit rate of correct retrieval from 300 candidates, given the associated brain sample (chance=0.3%); vice-versa for brain retrieval. Lite-Mind only uses a parameter quantity of 12.5M, achieving extremely high retrieval performance without using any model to close vectors of different modalities (see Appendix B.1 for remaining subject models).

| Method | Low-Level | | | | High-Level | | | |
|---|---|---|---|---|---|---|---|---|
| | PixCorr↑ | SSIM↑ | Alex(2)↑ | Alex(5)↑ | Incep↑ | CLIP↑ | Eff↓ | SwAV↓ |
| Lin et al. [20] | - | - | - | - | 78.2% | - | - | - |
| Takagi... [37] | - | - | 83.0% | 83.0% | 76.0% | 77.0% | - | - |
| Gu et al. [10] | .150 | .325 | - | - | - | - | .862 | .465 |
| Ozcelik... [26] | .254 | **.356** | 94.2% | 96.2% | 87.2% | 91.5% | .775 | .423 |
| BrainCLIP [22] | - | - | - | - | 86.7% | **94.8%** | - | - |
| MindEye [34] | **.309** | .323 | **94.7%** | **97.8%** | **93.8%** | 94.1 | **.645** | **.367** |
| MindEye-125M(LAION-5B) | .130 | .308 | 84.0% | 92.6% | 86.9% | 86.1% | .778 | .477 |
| Lite-Mind-8M(LAION-5B) | .125 | **.331** | 78.7% | 89.4% | **87.9%** | **88.7%** | **.724** | **.446** |

(b) LAION-5B retrieval performance. The upper part is based on various fMRI-image reconstruction methods using generative neural networks, while the lower part represents the retrieval performance of the large-scale dataset LAION-5B as an alternative to fMRI-image reconstruction. All results are averaged across four subjects, and Lite-Mind achieved higher performance with a lighter backbone, even exceeding the accuracy of many reconstruction methods.

**Table 1: Main results of Lite-Mind retrieval and LAION-5B retrieval on the NSD dataset.**

## 5.2 LAION-5B retrieval

Although the retrieval scale on NSD is large enough, we can still expand it to larger datasets, such as LAION-5B [33]. The final layer CLIP ViT-L/14 CLS embeddings for all 5 billion images are available at https://knn.laion.ai/ and can be queried for K-nearest neighbor lookup via the CLIP Retrieval client [2]. For LAION-5B retrieval, we train another similar **Lite-Mind-8M**, aligning voxels to the CLS embeddings of CLIP. For each test sample, we conduct a retrieval strategy in the same way as MindEye (first retrieve 16 candidate images using CLS embeddings, and the best image is selected based on having the highest cosine similarity to the fMRI voxel embeddings aligned to the final hidden layer of CLIP). Unlike MindEye, in the LAION-5B retrieval process, we used the Diffusion Projector to conduct image-to-image retrieval.

To compare the performance of large-scale dataset retrieval, we use metrics of image reconstruction for evaluation with previous image reconstruction methods. The experimental results are shown in Table 1b. The specific evaluation metrics in the table are as follows: PixCorr represents pixel-wise correlation between ground truth and retrieval/reconstruction images; SSIM is structural similarity index metric [39]; EfficientNetB1("Eff") [38] and SwAV-ResNet50("SwAV") [4] refer to average correlation distance; all other metrics refer to two-way identification (chance = 50%).

Experiments show that our DFT Backbone performs better than MindEye in high-level metrics, as shown in Table 1b, although the backbone only has 8M parameters (MindEye-CLS backbone still has

125M. Note that both MindEye and Lite-Mind perform LAION-5B retrieval based on CLS alignment, resulting in a smaller number of parameters compared to full models). The results indicate that an efficient DFT backbone can also assist downstream retrieval tasks. The retrieval visualization results of Lite-Mind on LAION-5B retrieval for Subject 1 are shown in Figure 4, and the specific retrieval performance and visualization results of each Subject are shown in Appendix B.2.

## 5.3 GOD zero-shot classification

Since fMRI data are much fewer and shorter than those of the NSD dataset usually, we conducted zero-shot classification tasks on a smaller fMRI dataset to verify Lite-Mind's generalization ability. The voxel lengths collected by Subjects on the GOD dataset range from 4133 to 4643, approximately 30% of NSD, and there are only 1200 fMRI-image pairs on the training dataset (4.8% of NSD).

We conduct zero-shot classification by using fMRI-to-image retrieval to find the specific image and using simple prompt text templates *"An image of [class]"* to obtain the category of retrieved images(CLIP has a classification accuracy of 76.2% on ImageNet without pre-training on it, so it can be seen as zero-shot classification). We train our **Lite-Mind-15.5M** to align voxels to images on the GOD dataset. We also replicated the MLP backbone and projector of MindEye, changing the residual block of $4096 \times 4096$ to $1024 \times 1024$ to accommodate the reduction of voxel length(We empirically found that the original $4096 \times 4096$ performed even

**Figure 4: Retrieval results of Lite-Mind from LAION-5B for Subject 1. The left column marked by a red box in every two columns represents the original image seen by the Subject, and the right column represents the image retrieved from LAION-5B.**

| Methods | Modality | Prompt | Subject1 | | Subject2 | | Subject3 | | Subject4 | | Subject5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | top-1 | top5 | top-1 | top5 | top-1 | top5 | top-1 | top5 | top-1 | top5 |
| CADA-VAE [32] | V&T | - | 6.31 | 35.70 | 6.45 | 40.12 | 17.74 | 54.34 | 12.17 | 36.64 | 7.45 | 35.04 |
| MVAE [40] | V&T | - | 5.77 | 31.51 | 5.40 | 38.46 | 17.11 | 52.46 | 14.02 | 40.90 | 7.89 | 34.63 |
| MMVAE [36] | V&T | - | 6.63 | 38.74 | 6.60 | 41.03 | 22.11 | 56.28 | 14.54 | 42.45 | 8.53 | 38.14 |
| MoPoE-VAE [23] | V&T | - | 8.54 | 44.05 | 8.34 | 48.11 | 22.68 | 61.82 | 14.57 | 58.51 | 10.45 | 46.40 |
| BraVL [7] | V&T | - | 9.11 | 46.80 | 8.91 | 48.86 | 24.00 | 62.06 | 15.08 | 60.00 | 12.86 | 47.94 |
| BrainClip-VAE [22] | V&T | Text | 8.00 | 42.00 | 24.00 | 52.00 | 20.00 | 58.00 | 20.00 | 58.00 | 20.00 | 46.00 |
| BrainClip-VAE [22] | V&T | CoOp | 14.00 | 54.00 | 20.00 | 60.00 | 21.33 | 64.67 | 16.67 | 66.67 | 18.00 | 52.00 |
| MindEye-211M [34] | V | Text | 6.00 | 40.00 | 20.00 | 60.00 | 28.00 | 70.00 | 14.00 | 54.00 | 10.00 | 64.00 |
| Lite-Mind-15.5M(ours) | V | Text | **26.00** | **74.00** | **28.00** | **70.00** | **30.00** | **80.00** | **34.00** | **76.00** | **24.00** | **66.00** |

**Table 2: Zero-shot visual stimulus classification on the GOD dataset. The test set contains 50 categories that have no overlapping with the training set (top-1 chance=2.0%). The results for CADA-VAE, MVAE, MMVAE, MoPoE-VAE, and BraVL are taken from [7]. MindEye(211M) is a smaller version consisting of MLP backbone and projectors, with a residual block size of $1024 \times 1024$ (See Appendix B.3 for retrieval performance on the GOD dataset).**

worse), and MindEye still has 210M parameters. As shown in Table 2, the experimental results indicate that our proposed Lite-Mind still guarantees robustness and establishes a new state-of-the-art for zero-shot classification on the GOD dataset without complex text prompt templates or auxiliary text representation training. On the contrary, due to the sudden decrease in dataset size and voxel length, the performance of MindEye significantly decreases, indicating excessive reliance on training data size for voxel value-wise mappings with large MLP Backbone. More retrieval and classification results are provided in Appendix B.3.

## 5.4 Ablations and visualization

We assess the effectiveness of different modules within Lite-Mind to investigate the lightweight and efficient performance in this session. All experiments below are for Subject 1 on the NSD dataset.

**Architectural Analysis.** We trained multiple DFT Backbone models with different depths to evaluate the impact of the number of layers of fMRI Spectrum Compression on retrieval efficiency(Table 3). As the Filter Block layers gradually deepen, the number of model parameters increases, the image retrieval accuracy gradually improves, and the improvement speed tends to be gradual. We interestingly found that the accuracy of brain retrieval tends to converge

faster than the image retrieval and only 6 layers of Filter Blocks are needed to approach the highest value. We have also conducted more ablation experiments on hyper-parameters to demonstrate the robustness of Lite-Mind. Refer to Appendix B.4 for it.

| Depth | FLOPs(G) | Parameters | Image Retrieval | Brain Retrieval |
|---|---|---|---|---|
| MindEye | 5.66 | 996M | 88.8% | 84.9% |
| 1 | 0.05 | 0.7M | 83.5% | 94.0% |
| 6 | 0.22 | 3.4M | 90.6% | 97.2% |
| 12 | 0.43 | 6.6M | 91.0% | 97.2% |
| 21 | 0.80 | 12.5M | 94.6% | 97.4% |

**Table 3: Ablation experiments on fMRI Spectrum Compression revealed a positive correlation between the depth of the Filter Blocks and enhanced forward retrieval accuracy.**

**Module Analysis.** We conducted ablation experiments on two main modules of DFT Backbone in Table 4. We completely removed the Filter blocks and aligned tokens directly to the image embeddings with frequency projector, achieving a forward retrieval accuracy of 65.8%, indicating the importance of frequency domain filtering for denoising and compression. Then we replaced FreMLP with the FC layer in the real domain, where tokens are flattened

697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812

into a tensor for alignment, similar to the last layer of the MLP backbone of MindEye. The model performance rapidly declined, indicating that tokens processed by DFT Backbone have better mapping performance with FreMLP in the frequency domain.

| Module | Image Retrieval | Brain Retrieval |
|---|---|---|
| DFT Backbone | 94.6% | 97.4% |
| w/o Filter Blocks | 65.8% | 74.7% |
| w/o FreMLP | 43.4% | 47.8% |

Table 4: Ablation experiments on modules of DFT Backbone.

**Embedding Dimension.** We conducted additional experiments to verify the impact of embedding dimensions on retrieval accuracy, as shown in Table 5. Among them, two different CLIP models (i.e., CLIP ViT/B-32 and CLIP ViT/L-14) were used to extract image embeddings, with a total of four embedding dimensions, by CLS embeddings and the last hidden layer, respectively. From the results, it can be seen that the retrieval accuracy of Lite-Mind is higher when the embedding dimensions of the image are longer and the representation of the image is richer. The results verify that the fine-grained alignment we mentioned above comes from the rich representation of the image, rather than the fMRI voxel value fully connected heavy MLP Backbone.

| Embeddings | Dimension | Parameters | Retrieval | |
|---|---|---|---|---|
| | | | Image↑ | Brain↑ |
| ViT-B/32 CLS | 512 | 6.7M | 57.5% | 61.8% |
| ViT-L/14 CLS | 768 | 6.7M | 60.3% | 64.2% |
| ViT-B/32 Hidden | 50×512 | 10.4M | 91.1% | 96.5% |
| ViT-L/14 Hidden | 257×768 | 12.5M | 94.6% | 97.4% |

Table 5: Retrieval performance with different CLIP embeddings for Subject 1 on the NSD dataset.

**Cerebral cortex.** To explore the impact of different cortical regions on retrieval accuracy, we used Takagi's [37] method to extract the *stream* ROI on the NSD dataset, covering the visual cortex of *nsdgenal*, and trained the individual retrieval DFT Backbones for visual cortices (*early*, *lateral*, *parietal*, *ventral*). As shown in Table 6, it can be observed that the early visual cortex has the greatest impact on retrieval accuracy. Although with a voxel length of 5917, Lite-Mind can still achieve retrieval accuracy of 85.0% and 93.4%, which is consistent with Takagi's research. They demonstrated that visual stimuli are almost dominated by the *early* visual cortex and indicate that *nsdgenal* still has a low signal-to-noise ratio, proving that fully connected MLP Backbones are unnecessary. Interestingly, for other cortical regions, both *lateral* and *ventral* have some impacts on retrieval accuracy, while *parietal* shows little.

**Visualization.** We visualize the CLS embeddings for LAION-5B using T-SNE in Figure 5. The figure shows that the diffusion projector successfully transformed the voxel CLS embeddings learned from contrastive learning into the image CLS embeddings, and the two kinds of embeddings are well fused for image-to-image retrieval on the LAION-5B dataset. We have also visualized the alignment process of voxel embeddings by contrastive learning on the NSD dataset in Appendix C.3 and learned filters of Filter Blocks in fMRI Spectrum Compression in Appendix C.5.
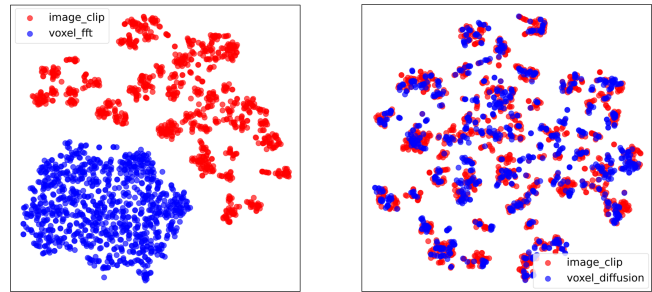


Figure 5: T-SNE of embeddings of 982 test fMRI by diffusion projector on LAION-5B retrieval for Subject 1. In the figure, image-clip represents image-clip CLS embeddings, while voxel-fft and voxel-diffusion represent voxel CLS embeddings by DFT Backbone or Diffusion Projector. The diffusion projector plays a role in bringing vectors closer, enabling image-to-image retrieval on the LAION-5B dataset.

| Region | Voxel Length | Parameters | Retrieval | |
|---|---|---|---|---|
| | | | Image↑ | Brain↑ |
| early | 5917 | 12.4M | 85.0% | 93.4% |
| lateral | 7799 | 12.4M | 26.9% | 29.0% |
| parietal | 3548 | 12.3M | 9.6% | 11.2% |
| ventral | 7604 | 12.3M | 18.1% | 22.4% |

Table 6: Retrieval performance with different cerebral cortex for Subject 1 on the NSD dataset.

## 6 LIMITATIONS AND FUTURE WORK

Due to inherent differences between Subjects, Lite-Mind still needs to train Subject-specific models. Although our exploratory experiments have shown that a common model can achieve comparable retrieval accuracy on four Subjects, it's still necessary to ensure sufficient training data. Further design beyond the model level is still needed for cross-Subject models, to achieve Subject-specific few-shot or even zero-shot. In addition, Lite-Mind performs worse on Subjects with shorter voxel lengths than on Subjects with longer, consistent with previous research, which lacks relevant neuroscience theoretical evidence. This provides a promising avenue for the interpretability of deep learning in neuroscience.

## 7 CONCLUSION

In this work, we propose Lite-Mind, an extremely lightweight brain representation learning paradigm based on Fourier Transform for cross-modal fMRI-to-image retrieval. Our DFT Backbone is an efficient means of obtaining fine-grained representation alignment between fMRI signals and visual stimuli. With high retrieval accuracy, Lite-Mind can transform many downstream fMRI tasks into retrieval-based tasks, such as zero-shot classification. Meanwhile, Lite-Mind's dependence on dataset size and voxel length is less than larger models, demonstrating excellent robustness across different fMRI datasets.

# REFERENCES

[1] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience* 25, 1 (2022), 116–126.

[2] Romain Beaumont. 2022. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them.

[3] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. 2020. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems* 33 (2020), 17766–17778.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems* 33 (2020), 9912–9924.

[5] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. 2023. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22710–22720.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[7] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. 2023. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[8] Max Ehrlich and Larry S Davis. 2019. Deep residual learning in the jpeg transform domain. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3484–3493.

[9] Chersoni Emmanuele, Santus Enrico, Huang Chu-Ren, Alessandro Lenci, et al. 2021. Decoding word embeddings with brain-based semantic features. *Computational Linguistics* 47, 3 (2021), 663–698.

[10] Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. 2022. Decoding natural image stimuli from fMRI data with a surface-based convolutional network. *arXiv preprint arXiv:2212.02409* (2022).

[11] Lionel Gueguen, Alex Sergeev, Ben Kadlec, Rosanne Liu, and Jason Yosinski. 2018. Faster neural networks straight from jpeg. *Advances in Neural Information Processing Systems* 31 (2018).

[12] John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. 2021. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587* (2021).

[13] Tomoyasu Horikawa and Yukiyasu Kamitani. 2017. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications* 8, 1 (2017), 15037.

[14] Yukiyasu Kamitani and Frank Tong. 2005. Decoding the visual and subjective contents of the human brain. *Nature neuroscience* 8, 5 (2005), 679–685.

[15] Viktor Kewenig, Christopher Edwards, Quitterie Lacome DEstalenx, Akilles Rechardt, Jeremy I Skipper, and Gabriella Vigliocco. 2023. Evidence of Human-Like Visual-Linguistic Integration in Multimodal Large Language Models During Predictive Language Processing. *arXiv preprint arXiv:2308.06035* (2023).

[16] Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. 2023. Brain-optimized inference improves reconstructions of fMRI brain activity. *arXiv preprint arXiv:2312.07705* (2023).

[17] Emirhan Koç and Aykut Koç. 2022. Fractional Fourier Transform in Time Series Prediction. *IEEE Signal Processing Letters* 29 (2022), 2542–2546.

[18] Henning Lange, Steven L Brunton, and J Nathan Kutz. 2021. From Fourier to Koopman: Spectral methods for long-term time series prediction. *The Journal of Machine Learning Research* 22, 1 (2021), 1881–1918.

[19] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824* (2021).

[20] Sikun Lin, Thomas Sprague, and Ambuj K Singh. 2022. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems* 35 (2022), 29624–29636.

[21] David Linden. 2021. Section 3 - Introduction. In *fMRI Neurofeedback*, Michelle Hampson (Ed.). Academic Press, 161–169. https://doi.org/10.1016/B978-0-12-822421-2.00008-9

[22] Yulong Liu, Yongqiang Ma, Wei Zhou, Guibo Zhu, and Nanning Zheng. 2023. BrainCLIP: Bridging Brain and Visual-Linguistic Representation via CLIP for Generic Natural Visual Stimulus Decoding from fMRI. *arXiv preprint arXiv:2302.12971* (2023).

[23] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[24] Weijian Mai and Zhijun Zhang. 2023. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv preprint arXiv:2308.07428* (2023).

[25] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. 2011. Encoding and decoding in fMRI. *Neuroimage* 56, 2 (2011), 400–410.

[26] Furkan Ozcelik and Rufin VanRullen. 2023. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. *arXiv preprint arXiv:2303.05334* (2023).

[27] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. 2020. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 11 (2020), 3833–3849.

[28] Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications* 9, 1 (2018), 963.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[31] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. 2021. Global filter networks for image classification. *Advances in neural information processing systems* 34 (2021), 980–993.

[32] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8247–8255.

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

[34] Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, et al. 2023. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. *arXiv preprint arXiv:2305.18274* (2023).

[35] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. 2019. Deep image reconstruction from human brain activity. *PLoS computational biology* 15, 1 (2019), e1006633.

[36] Yuge Shi, Brooks Paige, Philip Torr, et al. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems* 32 (2019).

[37] Yu Takagi and Shinji Nishimoto. 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.

[38] Mingxing Tan and V Le Quoc. [n. d.]. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020. *arXiv preprint arXiv:1905.11946* ([n. d.]).

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[40] Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems* 31 (2018).

[41] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. 2024. Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 8226–8235.

[42] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. 2020. Learning in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1740–1749.

[43] Ling Yang and Shenda Hong. 2022. Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion. In *International Conference on Machine Learning*. PMLR, 25038–25054.

[44] Yanchao Yang and Stefano Soatto. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4085–4095.

[45] Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. 2024. Frequency-domain MLPs are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems* 36 (2024).

[46] Shuxian Zou, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2022. Cross-modal cloze task: A new task to brain-to-word decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*. 648–657.