FLORA: A Unified Generalist Model for Visual Brain Decoding via Multimodal Neural Embeddings

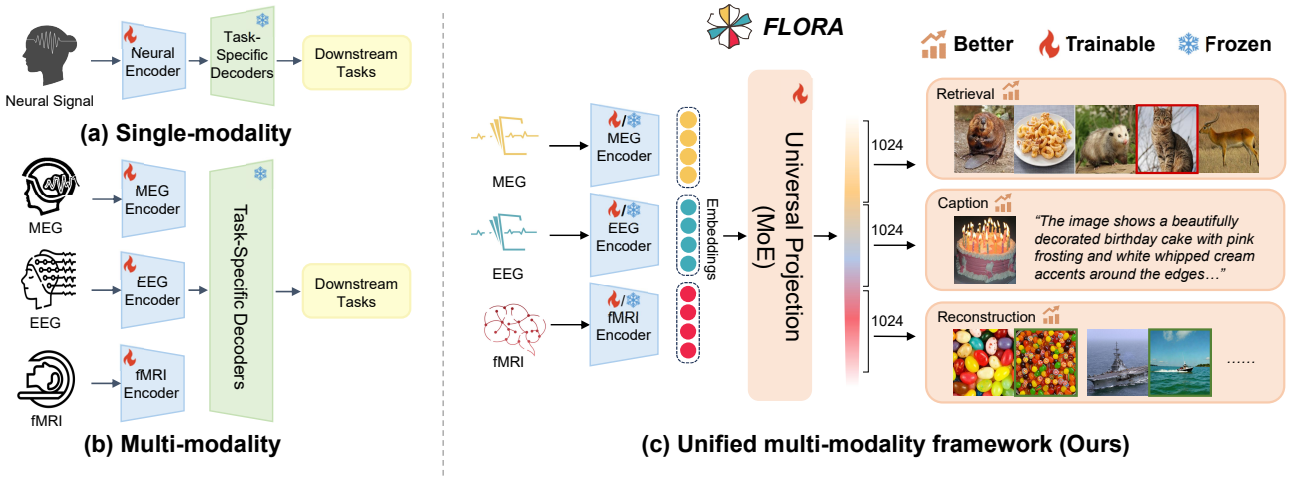


Figure 1. A comparative overview of visual decoding paradigms: (a) Single-modality decoding involves a neural encoder and a neural decoder, designed to process only one type of neural data at a time. (b) Multi-modality decoding employs distinct encoders tailored for various neural data types, with a unified decoder for subsequent tasks. (c) **Unified Multi-modality Model (Ours)**, featuring a shared, integrated encoding process that captures complementary representations across diverse neural data modalities. This architecture facilitates seamless interaction with multiple downstream tasks, embodying a **one-to-any** paradigm.

Abstract

Decoding visual information from neural data using artificial intelligence enhances our understanding of the human visual system. However, simultaneously acquiring paired neural data across modalities is challenging, leading most existing approaches to process these signals independently. This neglects their complementary characteristics and hinders decoding performance. In this study, we introduce **FLORA**, an end-to-end generalist model designed to integrate cross-modal neural data—including EEG, MEG, and fMRI—to construct a unified neural representation. FLORA employs multimodal large language models (MLLMs) alongside multimodal adapters and specialized diffusion model decoders, achieving superior performance on downstream tasks (e.g., neural signal retrieval and visual stimulus reconstruction) compared to single-modal approaches. By leveraging high-performance models, FLORA minimizes the number of parameters in the alignment and fusion layers, ensuring cost-effective fine-tuning. This design facilitates efficient training and seamless integration of extra modalities and datasets. Our approach holds promise for advancing our understanding of the brain's visual mechanisms and fostering new insights within the cognitive science and brain-computer interface communities. Our code is available at `https://anonymous.4open.science/r/FLORA-2C4A`.

1 Introduction

The brain processes external stimuli by receiving and encoding information through the coordinated activity of large-scale neural networks and intricate mechanisms. Gaining insight into these processes is essential, as it can reveal fundamental neural mechanisms and broaden potential applications in fields like brain-computer interfaces (BCIs). Recent advancements in computer science, particularly with generative models and pre-trained techniques, have driven substantial progress in decoding neural data from various neural modalities, including EEG (Li et al., 2024b; Kim et al., 2024), fMRI (Chen et al., 2023b; Caro et al., 2023), and MEG (Benchetrit et al., 2024; Cui et al., 2024). By leveraging the power of deep learning and unsupervised

algorithms, these methods are reshaping our capacity to interpret complex patterns of brain activity and enhancing the real-time decoding of neural signals.

One core challenge in achieving better decoding performance lies in establishing cross-modality representations for diverse neural data. Such data exhibit significant complexity and variability, influenced by factors like individual subjects, acquisition devices, visual content, and spatial-temporal dynamics. Existing approaches often rely solely on a single-modality (Pan et al., 2024; Xia et al., 2024c) or alignment among neural data, images, and text with contrastive learning (Fu et al., 2024; Jiang et al., 2024; Lin et al., 2022). These strategies either limit the decoding performance or introduce unconfirmed prior knowledge. Recent works develop multi-encoder structures such as NeuroBind (Yang et al., 2024a), which deploys distinct encoders tailored for specific types of neural data. Additionally, unified-encoder architectures like UMBRAE (Xia et al., 2024b) have aimed to establish a uniform representation across different neural data modalities. While these efforts demonstrate progress in enhancing performance for various downstream tasks, they remain constrained by the inherent limitations of single modalities and the lack of effective modal complementarity. This challenge highlights the critical need for more cohesive and innovative frameworks that not only integrate multiple modalities but also facilitate their unique characteristics to improve overall decoding properties.

Inspired by the power of multimodal large language models (MLLMs) in processing the input information in a wider variety of forms (Liu et al., 2024), we utilize it to integrate multi-modal neural data, for enhancing the encoding and decoding performance of complex brain activity in a previously unattainable way with single-modal data. In this work, we develop **FLORA**, a multimodal large language model that processes four modalities (i.e., EEG, MEG, fMRI, and behavioral data) within a unified framework. As illustrated in Fig. 1, FLORA comprises multimodal encoders, and a trainable mixture of experts (MoE) to perfrom universal projection across modalities. Unlike previous approaches, the encoder in FLORA is trained on a diverse range of multimodal data, including 16,540 paired EEG-image samples (Grootswagers et al., 2022), 19848 paired MEG-image samples, and 8540 paired fMRI-image samples (Hebart et al., 2023), ensuring shared representations across all modalities. To evaluate the FLORA model, we apply it for downstream tasks including retrieval, caption, and reconstruction compared to single-modality models. Through solid experiments, we find that FLORA outperforms mainstream single-modality-based decoder frameworks in all downstream tasks, benefiting from its unified training strategy. The unification in FLORA advances neuroscientific insights into visual representation in the brain and broadens practical Brain-Computer Interface (BCI) applications.

Our work has three main contributions.

- We propose a novel unified encoder to capture diverse neural data embeddings, realizing feature complementarity across heterogeneous neural modalities.
- Our framework, **FLORA**, enables seamless alignment and integration with various downstream tasks, significantly enhancing its adaptability and achieving SOTA performance across joint-subjects visual decoding benchmarks.
- To the best of our knowledge, **FLORA** is the first work to concurrently unify multiple modalities (EEG, MEG and fMRI) and task paradigms, providing a cohesive computational foundation for non-invasive neural decoding and BCI applications.

2 Related work

2.1 Visual decoding via uni-modal neural data

Significant progress has been made in visual decoding through single modalities such as EEG, fMRI, and MEG. Early research predominantly focused on leveraging deep learning techniques to classify (Zheng & Chen, 2021; Spampinato et al., 2019) and retrieve (Ye et al., 2022) visual stimuli based on neural recordings. In recent years, visual decoding has achieved unprecedented quality (Sun et al., 2023; Zhou et al., 2024; Benchetrit et al., 2024; Ma et al., 2024), driven by advancements in generative models (Ho et al., 2020; Rombach et al., 2022; Liu et al., 2022) and LLMs (Touvron et al., 2023; Brown et al., 2020). For instance, RealMind (Li et al., 2024a) utilizes multimodal models that combine the ability of semantic and geometric information, leading to improved decoding performance. Additionally, (Scotti et al., 2024b) demonstrates high-quality cross-subject reconstruction and retrieval capabilities, requiring minimal data, particularly valuable in scenarios with limited sample sizes. Moreover, (Benchetrit et al., 2024) successfully achieves real-time level retrieval and generation by capitalizing on the high temporal resolution of MEG, thereby enhancing the decoding of dynamic visual perception. These advancements underscore the practicality and effectiveness of utilizing artificial intelligence to gain insights into visual information processing and the functioning of the visual system.

2.2 Multimodal Large Models

Multimodal Large Language Models (MLLMs) have transformed AI by enabling seamless integration and generation of information across diverse modalities, including text, images, audio, video, and 3D objects. Foundational architectures like single-tower models (e.g., ViT (Dosovitskiy

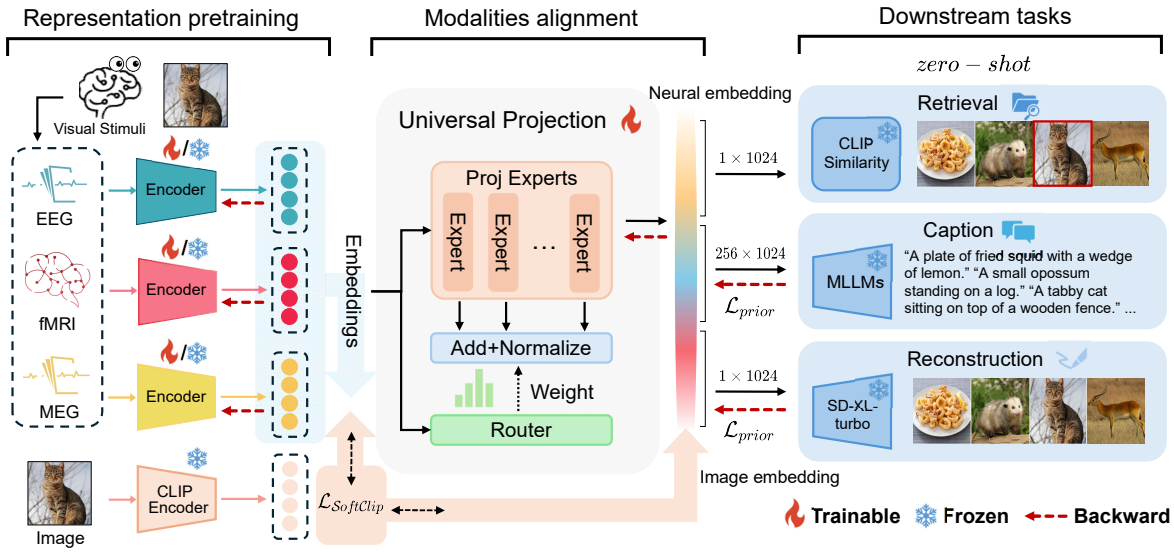


Figure 2. **Overall Architecture of FLORA.** FLORA comprises neural modality encoders, a universal projection, and a Mixture of Experts (MoE) projection module, with separate outputs for each modality. **Left:** The modality encoders, implemented as a time series backbone network with spatial-temporal convolutional layer, transforms the input signal into a sequence of tokens. The universal encoder is designed to align high-dimensional feature representations from heterogeneous data. **Middle:** The MoE projection module projects and aligns diverse neural modalities into a unified semantic representation space. **Right:** Various task heads facilitate different downstream tasks such as retrieval, captioning, and reconstruction. All modules are jointly trained during the same stage, optimizing computational efficiency.

et al., 2021)) and multi-tower models (e.g., CLIP (Radford et al., 2021)) pioneered image-text alignment, establishing the basis for multimodal learning. Building on these innovations, models like BLIP (Li et al., 2022), BLIP-2 (Li et al., 2023), Qwen-VL (Bai et al., 2023), and DALL·E (Ramesh et al., 2022) have significantly improved cross-modal understanding and generative capabilities. Inspired by the success of vision-language models, recent research has expanded to include audio (e.g., Tacotron (Shen et al., 2018), NaturalSpeech (Tan et al., 2022)), video (e.g., Create-to-Tell (Pan et al., 2018), DreamVideo (Wang et al., 2024a)), and 3D objects (e.g., DreamFusion (Poole et al., 2022)). Models like NExT-GPT (Wu et al., 2024) exemplify versatile multimodal capabilities, seamlessly interacting across text, images, audio, and video using distinct encoders tailored to various downstream tasks. Similarly, Meta-Transformer (Zhang et al., 2023) utilizes a unified parameter set to encode data from multiple modalities, effectively supporting a wide range of task types.

3 Method

Formulation. Let T represent the length of neural data, C the number of channels, and N the total number of data samples. Our objective is to derive neural embeddings $Z = \mathcal{E}(X) \in \mathbb{R}^{N \times F}$ from the brain activity data $X \in \mathbb{R}^{N \times C \times T}$, where $\mathcal{E}$ is the unified encoder and F is the projection dimension of the embeddings. Concurrently, we use the CLIP model to extract image embeddings $\hat{Z} \in \mathbb{R}^{N \times F}$ from images I. Our goal is to effectively align the multimodal representation with the image representation, as illustrated in Fig. 2.

3.1 Model Architecture

Our method aims to process the combination of multiple modalities of neural data via one unified model. All cross-modal neural features are efficiently aligned through the MoE module, maximizing complementary strengths across modalities. Multi-stage training strategies are applied to the encoding and alignment procedure. In the training phase, encoders are trained with neural data and image pairs using a contrastive learning framework. Then each modality is aligned by entering the *Unified Projector* based on each pre-trained encoder. In the inference phase, neural embeddings from the *Unified Projector* can be used for a variety of zero-shot tasks, including EEG/MEG/fMRI image classification, retrieval, and image reconstruction.

3.1.1 Neural Feature Extraction Module

This module is mainly employed to address the cross-subject challenges in multimodal neural data decoding as shown in Fig. 3. The core difficulty arises from the heterogeneous neural responses of different subjects to identical visual stimuli, leading to substantial inter-subject variability in

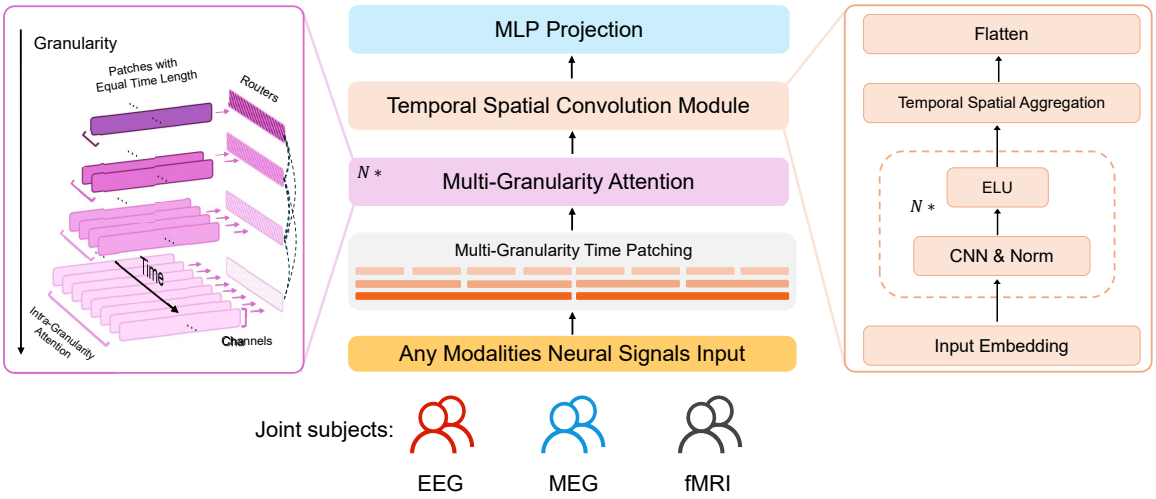


Figure 3. **Architecture of the Neural Encoder.** The original neural sequences from multiple variates are simultaneously embedded into tokens. Multi-granularity attention is applied to the embedded tokens of correlated variables, enhancing interpretability and revealing electrode-level correlations. The representations for each token are then assigned through the router layers. Subsequently, Temporal-Spatial convolution is employed to mitigate overfitting while improving the model's capacity for temporal-spatial representation learning.

the data. By incorporating multi-granular approaches, as proposed in (Wang et al., 2024c), into the encoder architecture, inter-subject heterogeneity can be mitigated, thereby facilitating the learning of shared joint embeddings across subjects.

For input neural signal $X \in \mathbb{R}^{T\times C}$, the architecture first performs *Multi-Granularity Time Patching* with exponentially increasing patch lengths $\{2^1, 2^2, ..., 2^n\}$: $x_p^{(i)} \in \mathbb{R}^{N_i\times(2^i\cdot C)}, N_i = \lceil T/2^i \rceil$. Then we get a group tokens $x^{(i)} \in \mathbb{R}^{N_i\times D}$ by:

$$x^{(i)} = x_p^{(i)} W^{(i)} + W_{pos}[1:N_i] + W_{gr}^{(i)} \tag{1}$$

where $W_{\text{pos}}[1:N_i] \in \mathbb{R}^{N_i\times D}$is the first N_i rows of the positional embedding W_{pos}, and a learnable granularity embedding $W_{\text{gr}}^{(i)} \in \mathbb{R}^{1\times D}$. And the router is used in the multi-granularity self-attention (as described later) for each granularity:

$$u^{(i)} = W_{pos}[N_i + 1] + W_{gr}^{(i)} \tag{2}$$

The embeddings undergo a two-stage *Multi-Granularity Attention* mechanism. In the first stage, intra-granularity attention captures temporal dependencies within each scale by concatenating patch embeddings with their corresponding router. The intermediate sequence of embeddings $z^{(i)} \in \mathbb{R}^{(N_i+1)\times D}$ is formed as:

$$z^{(i)} = \text{Concat}(x^{(i)}, u^{(i)}) \tag{3}$$

where $x^{(i)} \in \mathbb{R}^{N_i\times D}$ represents patch embeddings at scale i, and $u^{(i)} \in \mathbb{R}^{1\times D}$ is the router embedding. Self-attention is applied on each $z^{(i)}$ to update both $x^{(i)}$ and $u^{(i)}$. The collection of all scale-specific representations is:

$$Z = \text{Concat}(z^{(1)}, z^{(2)}, ..., z^{(n)}) \tag{4}$$

Then, inter-granularity attention facilitates information exchange across scales through router interactions. All router embeddings are aggregated:

$$U = \text{Concat}(u^{(1)}, u^{(2)}, ..., u^{(n)}) \in \mathbb{R}^{n\times D} \tag{5}$$

These router embeddings serve as query vectors to attend to features across scales. The attended features are concatenated as:

$$H_{\text{attn}} = \text{Concat}(x^{(1)}, x^{(2)}, ..., x^{(n)}) \in \mathbb{R}^{(\sum_{i=1}^{n} N_i)\times D} \tag{6}$$

Finally, H_{attn} is processed through a temporal-spatial convolution module to obtain H_{conv}, followed by an MLP to get the final encoding $Y \in \mathbb{R}^D$, which aligns with CLIP embeddings of the corresponding visual stimuli image.

3.1.2 Unified Projection Module

Aligning the multimodal neural representation space with the visual embedding space of the CLIP model is a crucial step, particularly given the complexity and heterogeneity of the multimodal neural imaging-image pairs. We employ a MoE module (Zhu et al., 2024), which has K projection experts $E_1, E_2, \ldots, E_K$, where each expert is a two-layer MLP, and a soft router R_{soft} to control the contribution of each expert.

Using a small network g, we use the soft router computes a score vector for each token, with a length equal to the

number of experts K. Finally, the Softmax function is applied to obtain the final routing vector, yielding the final routing vector. It can be formulated as:

$$R_{\text{soft}}(x_i) = \frac{Sigmoid(g(x_i))}{\sum Sigmoid(g(x_i))} \tag{7}$$

To alleviate the heterogeneity of different neural modalities. R_{soft} is a lightweight MLP designed to receive the inputs of different encoders, and calculate the routing weights $W_{\text{soft}} \in \mathbb{R}^{N\times F\times K}$ of each expert for each neural token, which can be formulated as:

$$W_{\text{soft}}(Z) = \sigma \cdot R_{\text{soft}}(Z) \tag{8}$$

where σ is the SoftMax function. Then we can obtain aligned neural tokens $Z \in \mathbb{R}^{N\times F}$ through a weighted sum of all experts' output as follows:

$$Z = \sum_{k=1}^{K} W_{\text{soft},k} \cdot E_k(Z) \tag{9}$$

where $w_{\text{soft},k}$ denotes the routing weight of the k-th projection expert.

3.2 Training Objective

We applied multiple approaches to loss functions, serving various downstream tasks. For retrieval tasks, we utilize CLIP Loss for contrastive learning(Radford et al., 2021). By aligning fused data F with corresponding image data I, FLORA could utilize potential of each modalities and achieve the identification of images seen by subjects with given neural data. Training consists of three progressive stages: visual-neural Data contrast training, visual-neural generation training, and supervised fine-tuning through **Universal Projection**.

3.2.1 High-level Visual Modeling

This pipline is primarily used to learn high level unified representation from the unified encoder. We use the contrastive learning functions inspired by (Scotti et al., 2024a) to train both the multi-tower encoder and *Unified Projection*:

$$\mathcal{L}_{\text{SoftCLIP}} = -\sum_{i,j=1}^{N} \frac{\exp\left(\frac{t_i\cdot t_j}{\tau}\right)}{Z_i} \cdot \log\left(\frac{\exp\left(\frac{p_i\cdot t_j}{\tau}\right)}{Z_i^{(p)}}\right) \tag{10}$$

$$Z_i = \sum_{m=1}^{N} \exp\left(\frac{t_i \cdot t_m}{\tau}\right), \quad Z_i^{(p)} = \sum_{m=1}^{N} \exp\left(\frac{p_i \cdot t_m}{\tau}\right) \tag{11}$$

Where $p_i = \mathcal{E}(x_i), t_i = \text{CLIP}_{\text{Image}}(y_i)$. For generative tasks, including captioning and reconstruction, we incorporate Mean Squared Error (MSE) loss to ensure consistent model training for regression-based tasks. As a result, the overall loss function of FLORA is formulated as a composite of these individual loss components. The optimization of model parameters θ is then governed by the following objective function:

$$\min_{\theta} \mathcal{L} = \mathcal{L}_{\text{BiMixCo|SoftCLIP}} + \alpha \cdot \mathcal{L}_{\text{MSE}} + \beta \cdot \mathcal{L}_{\text{prior}} \tag{12}$$

where α and β are hyperparameters to control the balance of each loss function.

3.2.2 Low-level Visual Modeling

Previous works (Scotti et al., 2024b; Li et al., 2024b; Zhang et al., 2024) have extensively used prior diffusion to enhance the transformation of neural data embeddings to image embedding distributions. We continue this approach with prior diffusion, but use the neural embeddings directly output by the *Unified Projection* module for retrieval evaluation—rather than after image prior enhancement, because this can more objectively measure the degree of alignment of neural data with the neural manifold of the image without over-reliance on image priors.

To train the low-level pipline, the original unified encoder goes through additional CNN upsampler to get $\hat{f}$, the approximation of original f. Then a new image $\hat{I}$ is reconstructed using the Stable Diffusion (Rombach et al., 2021) VAE decoder $D(\cdot)$ given $\hat{f}$, and a compound loss $\mathcal{L}$ is minimized:

$$\mathcal{L}_{\text{lowlevel}} = \|D(f)-D(\hat{f})\|+\|f-\hat{f}\|+\lambda_P\mathcal{L}_P(D(\hat{f})) \tag{13}$$

where $\mathcal{L}_P(\cdot)$ is a perceptual loss such as LPIPS (Zhang et al., 2018), and λ_P is loss weights.

4 Experiment

4.1 Dataset and Implementation

Dataset. We jointly trained FLORA with the THINGS-EEG2 (Gifford et al., 2022), THINGS-MEG, and THINGS-fMRI (Hebart et al., 2023) datasets to achieve a unified semantic representation. Notably, we noticed that relatively few decoding studies (Xue et al., 2024) have utilized the THINGS-EEG1 dataset (Grootswagers et al., 2022), which might be due to its low signal-to-noise ratio. This limitation is primarily attributed to two methodological factors: (i) the short presentation time of visual stimulus (50ms), and (ii) the lack of trial repetition (each stimulus viewed only once by each participant). Finally, we explored fine-tuning experiments using THINGS EEG1 in the Appendix B. The THINGS-EEG2 training set consists of 1654 concepts, each associated with 10 images, with 4 repetitions per image, while the testing set includes 200 concepts, each represented

Table 1. **Multimodal Retrieval Performance.** We present the retrieval performance of various methods on the THINGS-EEG2, THINGS-MEG and THINGS-fMRI datasets. Retrieval refers to the task of correctly identifying paired neural latent representations from the 200 or 100 class test set clip embeddings. Performance is evaluated across joint subject retrieval tasks, and we report results for several retrieval configurations: 2-way, 4-way, 10-way, as well as Top-1 and Top-5 accuracy for the 200-way retrieval task (100 for fMRI). FLORA is compared with previous SOTA methods, including NICE (Song et al., 2023), EEGNetV4 (Lawhern et al., 2018), B.D. (Benchetrit et al., 2024), MindEyeV2 (Scotti et al., 2024b), UMBRAE (Xia et al., 2024a), ATM-S (Li et al., 2024b), CongnitionCapturer (Zhang et al., 2024), MindBridge (Wang et al., 2024b), WAVE (Wang et al., 2024d) and MB2C (Wei et al., 2024).In subject results refers to Appendix F.

Model	THINGS-EEG2					THINGS-MEG					THINGS-fMRI				
	2-way Top-1	4-way Top-1	10-way Top-1	200-way Top-1	200-way Top-5	2-way Top-1	4-way Top-1	10-way Top-1	200-way Top-1	200-way Top-5	2-way Top-1	4-way Top-1	10-way Top-1	100-way Top-1	100-way Top-5
NICE	90.10	77.15	57.15	13.80	35.80	60.62	39.62	20.87	2.50	7.88	87.00	74.67	55.00	22.00	51.00
EEGNetV4	88.15	77.05	59.10	12.15	35.85	71.13	48.63	29.38	3.88	11.88	86.00	74.67	55.33	18.00	44.67
B.D.	90.05	76.85	55.85	12.45	34.15	62.12	35.50	17.63	1.88	6.50	72.67	58.33	38.00	6.67	26.00
MindEyeV2	90.25	79.85	64.30	17.35	43.65	64.25	40.25	20.37	2.13	8.38	92.00	80.33	64.33	21.33	60.33
UMBRAE	66.50	41.95	21.85	2.90	10.00	61.75	35.50	16.63	2.00	6.50	85.67	70.67	50.67	16.67	42.67
ATM-S	93.35	83.70	68.60	19.70	48.60	77.00	59.25	38.88	6.25	20.12	**92.67**	81.67	65.33	24.00	55.33
CongnitionCapturer	92.90	84.00	67.85	20.50	50.70	67.88	47.63	23.57	2.38	10.37	86.67	73.00	55.33	20.67	48.00
MindBridge	89.10	75.35	57.45	15.00	39.15	64.88	40.12	20.13	1.87	7.37	88.33	73.33	55.67	16.67	54.33
WAVE	89.40	75.85	58.80	15.00	39.05	67.50	44.75	24.00	3.00	11.38	88.67	71.67	52.33	20.33	45.33
MB2C	85.20	70.25	49.65	10.20	29.85	61.75	33.25	17.00	1.25	5.50	84.67	69.00	49.67	18.67	42.33
FLORA-unimodal	**95.55**	**86.90**	**73.45**	**25.35**	**57.30**	**81.75**	**64.50**	**46.62**	**8.00**	**24.38**	91.33	79.00	63.00	26.33	57.33
FLORA-multimodal	94.05	87.30	73.15	25.05	56.35	80.50	61.88	39.75	6.88	23.38	92.33	**84.67**	**70.67**	**28.33**	**63.33**

by a single image and 80 repeated trials. The original training set for THINGS-MEG contains 1854 concepts, each with 12 images (viewed once), and the testing set selects one unseen image from 200 of these concepts, repeated 12 times. The THINGS-fMRI training set comprises 720 concepts, each associated with 12 images (viewed once), and the testing set includes 100 concepts, with a single image selected from each and repeated 12 times. To facilitate zero-shot evaluation, we reclassified the data during experiments (Appendix D for more details).

Implementation and training setup. All experiments, including both training and inference procedures across modalities, were conducted using two NVIDIA A100-80GB GPUs. For visual encoding, we employed CLIP (ViT-L-14) to generate image embeddings. Each modality encoder was trained on the original training set of the THINGS dataset for 50 epochs, with a learning rate of 3e-4 and a batch size of 360, utilizing the AdamW optimizer. During the forward pass, batches from different modalities were processed sequentially, with each batch maintaining a size equal to the specified batch size. Hyperparameters were shared across all modalities.

4.2 Results on Retrieval

Multimodal retrieval Image retrieval metrics are used to quantify the extent of fine-grained image information captured in neural embeddings. The effectiveness of the image retrieval task is evaluated by computing the cosine similarity between embeddings derived from EEG, MEG, and fMRI signals, and the CLIP embeddings for a test set consisting of images. Tab. 1 presents the average retrieval performance across all subjects for EEG, MEG and fMRI.

4.3 Results on Reconstruction

Multimodal reconstruction Previous studies have made significant advancements in the task of reconstructing images from fMRI data (Takagi & Nishimoto, 2023; Chen et al., 2023a; Scotti et al., 2024a;b). The SDXL (Podell et al., 2023) + ip-adapter framework is utilized in the reconstruction pipline. Our approach builds upon the framework of MindEye2 (Scotti et al., 2024b), integrating state-of-the-art models such as SDXL (Podell et al., 2023) and FLUX for image reconstruction. What distinguishes our work is the ability to leverage the complementary strengths of the unified encoder, enabling the reconstruction of images from latent representations of multimodal neural data across different modalities using a single model. We employ a lightweight UNet as a prior diffusion denoising network to map the embeddings into the image space across different neural modalities. The cross-modal reconstruction performance results are presented in Tab. 2.

4.4 Results on Captioning

Multimodal caption Prior work (Li et al., 2024a) has demonstrated that directly reconstructing images from latent representations yields superior performance compared to first reconstructing the images and then generating captions. Building on this insight, we further employ Prior Diffusion to modulate the distribution of model features, thereby enhancing caption generation performance. Subsequently, we evaluated the captioning performance of our model by inputting the neural latents from the unified encoder to OpenFlamingo (Awadalla et al., 2023). The results are presented in Tab. 3.

Table 2. **Multimodal Visual Reconstruction.** Quantitative evaluation of the reconstruction quality for EEG, MEG and fMRI from all subjects. Comparative methods include MindEye (Scotti et al., 2024a), BrainDiffuser (Ozcelik & VanRullen, 2023), B.D. (Benchetrit Y, 2023), MindEye2(Scotti et al., 2024b), UMBRAE (Xia et al., 2024a) and ATM-S (Li et al., 2024b).It is noteworthy that, in contrast to other THINGS dataset-based works that conduct training and inference within single subjects, FLORA implements training and reconstruction in a cross-subject and cross-modality setting. For detailed descriptions of the evaluation details, please refer to the Appendix D.

Dataset	Method	Low-level			High-level			
		PixCorr ↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
NSD-fMRI	MindEye	0.130	0.308	0.917	0.974	0.936	0.942	0.369
	BrainDiffuser	0.254	0.356	0.942	0.962	0.872	0.915	0.423
	B.D.	0.305	0.366	0.962	0.977	0.910	0.917	0.410
	MindEye2	0.322	0.431	0.961	0.986	0.954	0.930	0.344
	UMBRAE	0.283	0.341	0.955	0.970	0.917	0.935	0.393
THINGS-fMRI	FLORA-multimodal	0.090	0.362	0.535	0.548	0.551	0.624	0.676
THINGS-MEG	B.D.	0.058	0.327	0.695	0.753	0.593	0.700	0.630
	B.D. (averaged)	0.076	0.336	0.736	0.826	0.671	0.767	0.584
	ATM-S	0.104	0.340	0.613	0.672	0.619	0.603	0.651
	FLORA-multimodal	0.070	0.291	0.553	0.613	0.567	0.619	0.663
THINGS-EEG2	ATM-S	0.160	0.345	0.776	0.866	0.734	0.786	0.582
	FLORA-multimodal	0.117	0.362	0.702	0.772	0.666	0.725	0.604

Table 3. **Multimodal Brain Captioning.** We compare the performance of various methods for brain captioning. To more clearly illustrate the superiority of our approach, the baseline method was trained and evaluated on the NSD dataset, while our method was trained and evaluated on the THINGS-fMRI dataset, in cross-subject settings.

Method	BLEU1 ↑	BLEU2 ↑	BLEU3 ↑	BLEU4 ↑	METEOR ↑	ROUGE ↑	CIDEr ↑	SPICE ↑	CLIP-S ↑	RefCLIP-S ↑
SDRecon (Takagi & Nishimoto, 2023)	36.21	17.11	7.72	3.43	10.03	25.13	13.83	5.02	61.07	66.36
OneLLM (Han et al., 2024)	47.04	26.97	15.49	9.51	13.55	35.05	22.99	6.26	54.80	61.28
UniBrain (Mai & Zhang, 2023)	-	-	-	-	16.90	22.20	-	-	-	-
BrainCap (Ferrante et al., 2023)	55.96	36.21	22.70	14.51	16.68	40.69	41.30	9.06	64.31	69.90
UMBRAE (Xia et al., 2024a)	57.84	38.43	25.41	17.17	18.70	42.14	53.87	12.27	66.10	72.33
FLORA-unimodal w/o prior fMRI*	36.71	30.58	25.36	18.80	41.14	40.52	4.46	16.19	17.55	73.37
FLORA-unimodal w/ prior fMRI*	33.73	27.77	22.41	16.05	41.73	39.21	4.74	15.05	17.71	59.81

* The evaluation metrics for FLORA were computed using the **nltk** evaluation framework.

4.5 In-depth Analysis

Visual Concept Localization We adopted a methodology similar to that presented in (Shen et al., 2024) for localizing semantic concepts within neural signals. Specifically, we employed Honeybee (Cha et al., 2024) to extract the target concepts from natural language with the prompt "Describe the main concept in the picture in three words". These concepts were encoded by the CLIP ViT-L-14 text encoder and used as target representations for integrated Grad-CAM, facilitating the spatial localization of these concepts within brain signals. We trained a unified encoder incorporating all three input modalities. The final layers of this model were leveraged to extract semantic features. Fig. 4 visualizes the localization outcomes on the THINGS-fMRI dataset, showing the discrimination of various semantic information within brain signals in response to novel visual stimuli.

We conducted ablation studies on the localized semantic concepts. After identifying the relevant concepts within the original fMRI voxels from visual cortex, we selectively masked the corresponding voxels. Feature extraction and visual reconstruction were subsequently performed using these modified brain signals. In Fig. 5, the targeted removal of fMRI voxels from specific brain regions associated with particular semantic concepts led to the exclusion of these semantics in the reconstructed visual representation.

5 Discussion and Conclusion

Here, we present FLORA, a unified generalist model that employs a shared encoder to extract cohesive neural representations across three distinct modalities: EEG, MEG, and fMRI. To the best of our knowledge, this represents the first work to align visual representations across multiple neural modalities using a unified encoder. FLORA achieves substantial performance gains across a range of downstream neural decoding tasks, representing a significant step forward in the development of a unified framework for cross-modal visual neural decoding.

Our framework has several limitations that warrant further investigation. First, the samples from different neural data modalities are imbalanced. Such imbalanced distri-

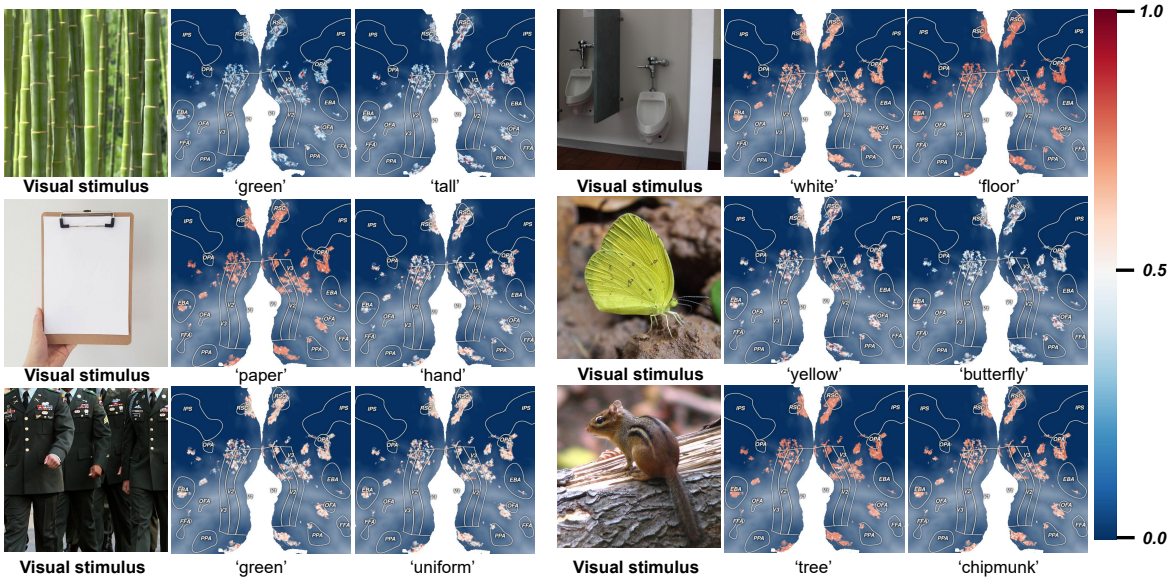


Figure 4. Gradient heatmaps of brain activity generated by FLORA. It shows the neural representation of different semantic information in the brain in response to the same visual stimulus, such as the semantics 'green' and 'tall' from the bamboo image.

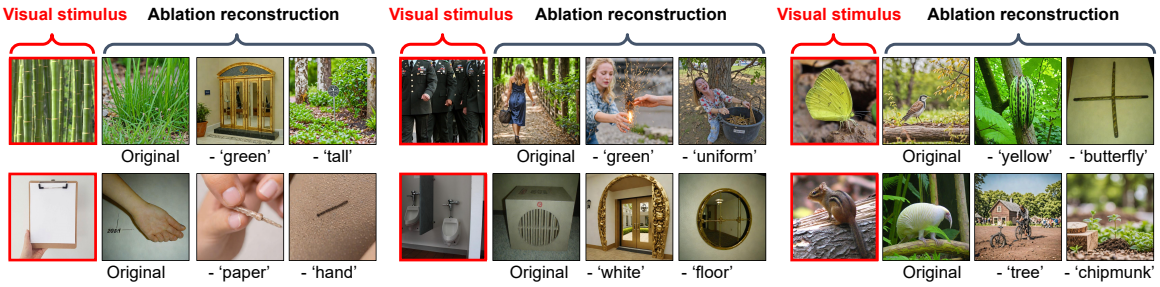


Figure 5. Concept analysis on THINGS-fMRI using FLORA through semantic signal nullification. We show the impacts on visual reconstruction by ablating certain concepts from visual stimuli.

butions of samples across modalities and classes pose a challenge for the unified encoder in effectively capturing modality-specific representations, thereby constraining the overall performance of the framework. To mitigate this issue, one potential solution could involve pre-training the modality-specific encoders on larger and more balanced neural datasets, which may enhance their ability to generalize across modalities. Second, due to the shared parameters and representation spaces, our framework underperforms compared to some SOTA methods in certain single-modality downstream tasks, such as reconstruction and caption generation. This performance gap could potentially be addressed by leveraging larger datasets and more sophisticated model architectures. Finally, inherent limitations of THINGS series datasets, including their limited size and repetitive nature, restrict FLORA's ability to enforce consistent category distributions between training and test sets across modalities. For example, the fMRI training and test sets exhibit an inclusion relationship in their category sets, which may introduce biases. To overcome these challenges, larger-scale, cross-modal datasets with more consistent category distributions are of high need.

Despite these limitations, FLORA demonstrates the potential of multimodal grand unified neural decoding by effectively leveraging the complementary information from different modalities to enhance performance across various tasks. Future work may prioritize two key directions: (i) the utilization of larger neural datasets, both paired and unpaired, to train robust multimodal foundation models, and (ii) the exploration of cutting-edge unified model architectures that incorporate advanced components for representation, alignment, and fusion stages, tailored to diverse neural decoding tasks. With such a unified framework, beyond the visual concept analysis (Fig. 4), more scientific questions can be studied, such as how these semantic representations across modalities are unified and adapted to personal experiences and knowledge.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Benchetrit, Y., Banville, H., and King, J.-R. Brain decoding: toward real-time reconstruction of visual perception. In *The Twelfth International Conference on Learning Representations*, 2024.

Benchetrit Y, Banville H, K. J. R. Brain decoding: toward real-time reconstruction of visual perception. *arXiv*, 2310: 19812, 2023.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Caro, J. O., de O. Fonseca, A. H., Averill, C., Rizvi, S. A., Rosati, M., Cross, J. L., Mittal, P., Zappala, E., Levine, D., Dhodapkar, R. M., Abdallah, C. G., and van Dijk, D. Brainlm: A foundation model for brain activity recordings. *bioRxiv*, 2023. doi: 10.1101/2023.09.12.557460. URL https://www.biorxiv.org/content/early/2023/09/13/2023.09.12.557460.

Cha, J., Kang, W., Mun, J., and Roh, B. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13817–13827, 2024.

Chen, Z., Qing, J., Xiang, T., Yue, W. L., and Zhou, J. H. 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*, pp. 22710–22720, 2023a.

Chen, Z., Qing, J., Xiang, T., Yue, W. L., and Zhou, J. H. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding, 2023b. URL https://arxiv.org/abs/2211.06956.

Cui, W., Jeong, W., Thölke, P., Medani, T., Jerbi, K., Joshi, A. A., and Leahy, R. M. Neuro-gpt: Towards a foundation model for eeg, 2024. URL https://arxiv.org/abs/2311.03764.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Ferrante, M., Ozcelik, F., Boccato, T., VanRullen, R., and Toschi, N. Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*, 2023.

Ferrante, M., Boccato, T., and Toschi, N. Towards neural foundation models for vision: Aligning eeg, meg and fmri representations to perform decoding, encoding and modality conversion. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

Fu, H., Shen, Z., Chin, J. J., and Wang, H. Brainvis: Exploring the bridge between brain and visual signals via image reconstruction, 2024. URL https://arxiv.org/abs/2312.14871.

Gifford, A. T., Dwivedi, K., Roig, G., and Cichy, R. M. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Grootswagers, T., Zhou, I., Robinson, A. K., Hebart, M. N., and Carlson, T. A. Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Sci Data*, 9(1):3, January 2022. ISSN 2052-4463. doi: 10.1038/s41597-021-01102-7. URL https://www.nature.com/articles/s41597-021-01102-7.

Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., and Yue, X. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., and Baker, C. I. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Jiang, S., Meng, Z., Liu, D., Li, H., Su, F., and Zhao, Z. Mindshot: Brain decoding framework using only one image, 2024. URL https://arxiv.org/abs/2405.15278.

Kim, J. W., Alaa, A., and Bernardo, D. Eeg-gpt: Exploring capabilities of large language models for eeg classification and interpretation, 2024. URL https://arxiv.org/abs/2401.18006.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., et al. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

Li, D., Qin, H., Wu, M., Cao, Y., Wei, C., and Liu, Q. Realmind: Zero-shot eeg-based visual decoding and captioning using multi-modal models, 2024a. URL https://arxiv.org/abs/2410.23754.

Li, D., Wei, C., Li, S., Zou, J., and Liu, Q. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024b.

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

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730–19742, 2023.

Lin, S., Sprague, T., and Singh, A. K. Mind reader: Reconstructing complex images from brain activities, 2022. URL https://arxiv.org/abs/2210.01769.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Liu, Z., Dong, Y., Liu, Z., Hu, W., Lu, J., and Rao, Y. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution, 2024. URL https://arxiv.org/abs/2409.12961.

Ma, S., Wang, L., Hou, S., and Yan, B. Aligned with llm: a new multi-modal training paradigm for encoding fmri activity in visual cortex, 2024. URL https://arxiv.org/abs/2401.03851.

Mai, W. and Zhang, Z. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv preprint arXiv:2308.07428*, 2023.

Ozcelik, F. and VanRullen, R. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

Pan, H., Li, Z., Fu, Y., Qin, X., and Hu, J. Reconstructing visual stimulus images from eeg signals based on deep visual representation model, 2024. URL https://arxiv.org/abs/2403.06532.

Pan, Y., Qiu, Z., Yao, T., Li, H., and Mei, T. To create what you tell: Generating videos from captions, 2018. URL https://arxiv.org/abs/1804.08264.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion, 2022. URL https://arxiv.org/abs/2209.14988.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Scotti, P., Banerjee, A., Goode, J., et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36, 2024a.

Scotti, P. S., Tripathy, M., Torrico, C., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K. A., et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *Forty-first International Conference on Machine Learning*, 2024b.

Shen, G., Zhao, D., He, X., Feng, L., Dong, Y., Wang, J., Zhang, Q., and Zeng, Y. Neuro-vision to language:

Image reconstruction and interaction via non-invasive brain recordings. *arXiv preprint arXiv:2404.19438*, 2024.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, 2018. URL https://arxiv.org/abs/1712.05884.

Song, Y., Liu, B., Li, X., Shi, N., Wang, Y., and Gao, X. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*, 2023.

Spampinato, C., Palazzo, S., Kavasidis, I., Giordano, D., Shah, M., and Souly, N. Deep learning human mind for automated visual classification, 2019. URL https://arxiv.org/abs/1609.00344.

Sun, J., Li, M., Chen, Z., Zhang, Y., Wang, S., and Moens, M.-F. Contrast, attend and diffuse to decode high-resolution images from brain activities, 2023. URL https://arxiv.org/abs/2305.17214.

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

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Soong, F., Qin, T., Zhao, S., and Liu, T.-Y. Naturalspeech: End-to-end text to speech synthesis with human-level quality, 2022. URL https://arxiv.org/abs/2205.04421.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Wang, C., Gu, J., Hu, P., Xu, S., Xu, H., and Liang, X. Dreamvideo: High-fidelity image-to-video generation with image retention and text guidance, 2024a. URL https://arxiv.org/abs/2312.03018.

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

Wang, Y., Huang, N., Li, T., Yan, Y., and Zhang, X. Medformer: A multi-granularity patching transformer for medical time-series classification. *arXiv preprint arXiv:2405.19363*, 2024c.

Wang, Y., Turnbull, A., Xiang, T., Xu, Y., Zhou, S., Masoud, A., Azizi, S., Lin, F. V., and Adeli, E. Decoding visual experience and mapping semantics through whole-brain analysis using fmri foundation models. *arXiv preprint arXiv:2411.07121*, 2024d.

Wei, Y., Cao, L., Li, H., and Dong, Y. Mb2c: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8992–9000, 2024.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

Xia, W., de Charette, R., Öztireli, C., and Xue, J.-H. Umbrae: Unified multimodal decoding of brain signals. *arXiv preprint arXiv:2404.07202*, 2024a.

Xia, W., de Charette, R., Oztireli, C., and Xue, J.-H. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, pp. 242–259. Springer, 2024b.

Xia, W., de Charette, R., Öztireli, C., and Xue, J.-H. Dream: Visual decoding from reversing human visual system, 2024c. URL https://arxiv.org/abs/2310.02265.

Xue, S., Jin, B., Jiang, J., Guo, L., and Liu, J. A hybrid local-global neural network for visual classification using raw eeg signals. *Scientific Reports*, 14(1):27170, 2024.

Yang, F., Feng, C., Wang, D., Wang, T., Zeng, Z., Xu, Z., Park, H., Ji, P., Zhao, H., Li, Y., and Wong, A. Neurobind: Towards unified multimodal representations for neural signals, 2024a. URL https://arxiv.org/abs/2407.14020.

Yang, F., Feng, C., Wang, D., Wang, T., Zeng, Z., Xu, Z., Park, H., Ji, P., Zhao, H., Li, Y., et al. Neurobind: Towards unified multimodal representations for neural signals. *arXiv preprint arXiv:2407.14020*, 2024b.

Ye, Z., Yao, L., Zhang, Y., and Gustin, S. See what you see: Self-supervised cross-modal retrieval of visual stimuli from brain activity, 2022. URL https://arxiv.org/abs/2208.03666.

Zhang, K., He, L., Jiang, X., Lu, W., Wang, D., and Gao, X. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. *arXiv preprint arXiv:2412.10489*, 2024.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pp. 586–595, 2018.

Zhang, Y., Gong, K., Zhang, K., Li, H., Qiao, Y., Ouyang, W., and Yue, X. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023.

Zheng, X. and Chen, W. An attention-based bi-lstm method for visual object classification via eeg. *Biomedical Signal Processing and Control*, 63:102174, 2021. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2020.102174. URL https://www.sciencedirect.com/science/article/pii/S174680942030313X.

Zhou, Q., Du, C., Wang, S., and He, H. Clip-mused: Clip-guided multi-subject visual neural information semantic decoding, 2024. URL https://arxiv.org/abs/2402.08994.

Zhu, X., Hu, Y., Mo, F., Li, M., and Wu, J. Unimed: A unified medical generalist foundation model for multi-task learning via connector-moe. *arXiv preprint arXiv:2409.17508*, 2024.

A Appendix Overview

- **Sec. B:** Additional Ablation Experiments.
- **Sec. C:** Additional Implementation Details.
- **Sec. D:** Evaluation Details.
- **Sec. E:** Comparison with Prior Works.
- **Sec. F:** Additional Qualitative Results.

B Additional Ablation Experiments

B.1 Ablation Study of FLORA

To assess the contribution of each component within our framework, we conducted ablation studies on various model modules in Tab. 4. Using Medformer (Wang et al., 2024c) as the baseline, we evaluated the impact of the temporal-spatial convolution module and the Mixture of Experts (MoE) module. Additionally, we explored replacing the MoE module with the Perceiver as the universal projection module; however, this approach resulted in a prohibitively large number of parameters, slower computation times, and suboptimal performance.

The experimental results substantiate that our proposed architecture effectively optimizes the trade-off between model capacity and computational efficiency. The temporal-spatial convolutional module preserves predictive performance while substantially reducing the parameter count, thereby facilitating the scalable processing of high-dimensional neural datasets. More critically, our soft-routing universal projection mechanism, when trained with heterogeneous multimodal neural data, learns invariant and transferable feature representations. Traditional MoE module selects a single expert based on the maximum score, while softrouter implements a weighted combination of experts. This architectural paradigm is particularly advantageous for joint-subject modeling, as the exposure to diverse neurophysiological patterns across modalities during training enhances FLORA's robustness in addressing inter-subject variability while maintaining a significantly lower computational burden compared to alternatives such as Perceiver (Alayrac et al., 2022). The empirical results demonstrate that this routing strategy in universal projection facilitates knowledge transfer across modalities, contributing to strong generalization capabilities in zero-shot cross-modal inference.

The ablation study corroborates our design rationale, highlighting the efficacy of multimodal training in bolstering model generalization—an attribute that becomes increasingly pivotal when scaling to expansive neural datasets encompassing extensive subject pools and multiple trial repetitions.

B.2 Supervised Scaling Up on Data Size

We analyzed the effect of training set size on joint subject training performance using THINGS-EEG2, as illustrated in 6A. This factor is critical for model efficacy. The training set comprised 1654 distinct concepts $\times$ 10 image conditions, each repeated four times. We preserved all four repetitions for training. Expanding the dataset size by increasing the number of conditions yielded substantial improvements in decoding accuracy, particularly in transitions from 20% to 40% and from 40% to 60%. Additionally, augmenting repetitions contributed significantly to performance gains, notably from 80% to 100%. These findings indicate that, under joint-subject training, larger datasets facilitate further performance enhancements. Our approach demonstrates strong scalability, underscoring its potential for improving neural decoding with increased data availability.

A key advantage of our cross-modality joint-subject training framework is its capacity to enable subject adaptation with minimal training data. To rigorously assess the generalization capability of our approach, we leverage the pretrained unified encoder on THINGS-EEG2, THINGS-MEG, and THINGS-fMRI to adapt the joint-subject model to THINGS-EEG1, utilizing varying amounts of training data, as illustrated in Fig. 6B. The dataset partitioning strategy for training and evaluation in THINGS-EEG1 remains consistent with its original split. We perform adaptation evaluation on a 200-way image retrieval task (chance level: 0.5%). Given

Table 4. Retrieval performance of FLORA under varying input projection mechanisms. The first two rows compare time-series Transformers trained using an attention-based algorithm without any projection or feature fusion. The following rows illustrate the impact of FLORA enhancements. "Medformer": A multi-granularity time-series model designed for medical data. "Medformer+TSConv": Medformer with the addition of temporal-spatial convolution. "Re. EEG": Retrieval performance on the THINGS-EEG2 dataset. "Re. MEG": Retrieval performance on the THINGS-MEG dataset. "Re. fMRI": Retrieval performance on the THINGS-fMRI dataset.

#	Description	Para.	Model	Uni. Proj.	Softrouter	Re. EEG ↑	Re. MEG ↑	Re. fMRI ↑
1	Med. (Linear)	97.9M	Time series	✗	✗	18.15	3.50	25.00
2	Med. (TSConv)	18.67M	FLORA-uni.	✗	✗	26.50	3.00	24.00
3	+Uni. Proj.	162.0M	FLORA-multi.	✓	✗	26.45	8.38	24.33
4	+Softrouter	162.0M	FLORA-uni.	✓	✓	25.35	8.00	26.33
	+Softrouter	162.0M	FLORA-multi.	✓	✓	25.05	6.88	28.33

the inherently low signal-to-noise ratio in EEG data, the zero-shot performance of our model approximates chance level. However, after 50 epochs of fine-tuning, the model exhibits notable scalability, underscoring its potential for efficient subject adaptation and robust cross-modality transfer.

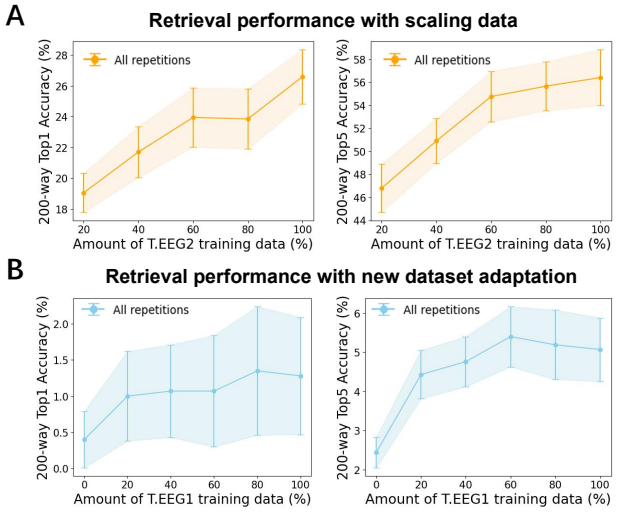


Figure 6. Effect of data size in training and test set (THINGS-EEG2), and weakly-supervised subject adaptation (THINGS-EEG1).

C Additional Implementation Details

C.1 Neural Modality Encoders

EEG Encoder. The shape of an EEG signal is $R^{63\times250}$. With Medformer, we get multiple sets of tokens output $x \in \mathbb{R}^{221\times250}$. By temporal-spatial convolution layers, we then resize the tensor $x \in \mathbb{R}^{221\times250}$ into a 1D tensor $x \in \mathbb{R}^{1\times1024}$.

MEG Encoder. The shape of an MEG signal is $R^{271\times201}$. With Medformer, we get multiple sets of tokens output $x \in \mathbb{R}^{178\times250}$. By temporal-spatial convolution layers, we then resize the tensor $x \in \mathbb{R}^{178\times250}$ into a 1D tensor $x \in \mathbb{R}^{1\times1024}$.

fMRI Encoder. In the THINGS-fMRI data set, the original voxel number of visual ROI (Region of Interest) of each of the three subjects was 6036, 5944 and 5238. In order to facilitate model processing, pad was unified to 7000 when data was loaded. So the shape of an fMRI signal is R^{7000}. We tokenize it with a linear layer: Linear($C_{\text{in}} = 7000, C_{\text{out}} = 8192$). We then resize the output tensor $x \in \mathbb{R}^{8192}$ into a 2D tensor $x \in \mathbb{R}^{8\times1024}$ to align with the input of the transformer encoder. With Medformer, we get multiple sets of tokens output $x \in \mathbb{R}^{899\times250}$. By temporal-spatial convolution layers, we then resize the tensor $x \in \mathbb{R}^{899\times250}$ into a 1D tensor $x \in \mathbb{R}^{1\times1024}$.

C.2 Low-Level Pipeline

Compared to vision-centric pretraining paradigms such as ViT, ResNet, and DINO, the CLIP vision model exhibits a deficiency in capturing fine-grained low-level visual features. To mitigate this limitation, our framework integrates a dedicated low-level visual reconstruction pipeline. Specifically, we aim to reconstruct fundamental perceptual attributes—such as contour, posture, and orientation—by leveraging EEG-derived representations. This is achieved through an alignment mechanism with the latent space of a variational autoencoder (VAE), facilitating the recovery of pixel-level structural information. By enforcing this latent alignment, our approach enhances the model's capacity to infer and preserve crucial low-level visual semantics from neural signals, thereby improving the fidelity of EEG-to-vision mappings.

We trained the low-level reconstruction pipeline for 200 epochs, employing a latent mean squared error (MSE) loss in conjunction with a contrastive learning loss and a variational autoencoder (VAE) image reconstruction loss. The objective was to align the $4 \times 64 \times 64$ EEG-derived latent representations—obtained via a projection layer and an up-

sampling CNN—with the VAE latent space. However, we observed that the reconstruction loss and contrastive learning loss underperformed compared to solely optimizing the latent space loss, while also imposing significantly higher GPU memory demands.

Furthermore, our investigation revealed that incorporating a low-level visual model for knowledge distillation in the low-level reconstruction pipeline not only failed to enhance VAE latent alignment but also exacerbated overfitting. These findings suggest that zero-shot low-level visual reconstruction from EEG signals lacks stability and may introduce misleading artifacts in the generated outputs.

In our framework, when utilizing the low-level reconstruction pipeline, we typically set the inference steps of SDXL to 10 (or SDXL-Turbo to 4) and configure the image-to-image denoising strength to 0.5.

D Evaluation Details

In this section, we will give more evaluation details.

Visual Retrieval tasks. We use forward retrieval (all image embeddings in test set is retrieved using one neural data embedding) for evaluation. For the neural data of each modality, we output it as $x \in \mathbb{R}^{1\times1024}$ and calculate the dot product similarity with the image embedding, taking the image with the greatest similarity as the image output by zero-shot. For THINGS-EEG1, the test set has 12 conditions per concept, so we only take the first image condition among the 200 concepts in the test set, that is, we use 200 images from 200 categories to test. For THINGS-EEG2, the test set has only 1 condition per concept, so we test with 200 images from 200 categories. For THINGS-MEG, there are 200 concepts in the test set, each of which has 12 conditions, so we only use the first condition of each concept as the test, with a total of 200 images from 200 categories to test. For THINGS-fMRI, the test set is 100 concepts, where each concept has 1 condition, so a total of 200 images from 200 categories are tested.

Visual Reconstruction tasks. The test set configuration of the visual reconstruction task is exactly the same as that of the retrieval task, except that the image reconstruction has more indicators to measure the performance of the image reconstruction. It should be emphasized that, in contrast to existing approaches on the THINGS dataset which train models independently for each subject's neural data and conduct visual reconstruction, FLORA-multimodal necessitates addressing highly heterogeneous multimodal and cross-subject neural data in both the training phase and visual reconstruction inference evaluation. Moreover, within the fMRI modality, FLORA employs the THINGS-fMRI dataset, which is smaller in scale than the NSD dataset used in existing major works, thus requiring the model to more efficiently utilize limited data resources.

Visual Captioning tasks. The idea of employing Prior diffusion to alter the distribution of model features is inspired by the two-stage image reconstruction approach described in ATM-S (Li et al., 2024b), with the expectation of achieving similar effects in the domain of caption generation. Although the final quantitative results show only marginal differences, the incorporation of Prior diffusion leads to more structurally coherent and semantically accurate captions that better reflect the content of the original images. The test set configuration for the visual captioning task is exactly the same as for the retrieval task. The THINGS image data set lacks the original text captions annotation. To generate captions for the test set images as ground truth, we used the prompt "Please provide a concise and formal sentence to describe the image, beginning with 'An image of,' and keeping it to approximately 10 words." from the large model Kimi to obtain descriptions for each image in the test set. By using different MMLMs as interfaces, we can generate captions of different qualities from the latent, and we give an assessment of captioning quality for different MMLMs adapted to FLORA.

E Comparison with Prior Works

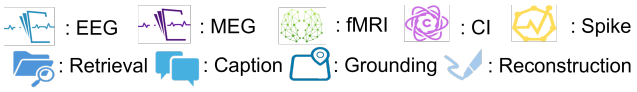
The key distinction between FLORA and previous neural decoding approaches lies in its demonstration that specialized neural encoders, when coupled with a universal projection layer, can effectively align diverse neural modalities with visual representations in a unified framework. As evidenced in Tab. 5, while prior works typically develop independent architectures for each modality, FLORA achieves multimodal integration across EEG, MEG, and fMRI signals while maintaining a compact parameter footprint (7.36M). The empirical results (Tab. 1) demonstrate competitive performance across modalities compared to modality-specific models, while ablation studies (Tab. 4) reveal that joint training within this unified framework provides particular benefits for modalities with limited data availability.

The effectiveness of FLORA's architecture in handling diverse neural signals suggests broader implications for scalable neural decoding. By successfully integrating multiple modalities through a shared projection space while maintaining computational efficiency, FLORA provides insights into architectural design principles for larger-scale neural models. The framework's demonstrated capabilities in addressing fundamental challenges - from cross-subject variability to zero-shot generalization - while preserving modality-specific characteristics, establish a foundation for developing more comprehensive neural decoding systems. These results not only validate the feasibility of unified multimodal frameworks but also suggest promising directions

for scaling neural decoding models toward more ambitious multimodal applications.

Table 5. Comparison between FLORA and related works on visual retrieval, reconstruction, and caption tasks. Methods compared include MindEye2 (Scotti et al., 2024b), NICE (Song et al., 2023), B.D. (Benchetrit et al., 2024), N.F. (Ferrante et al., 2024), NeuroVL (Shen et al., 2024), UMBRAE (Xia et al., 2024b), V.D. (Li et al., 2024b), and NeuroBind (Yang et al., 2024b).

Method	Modalities	Tasks
MindEye2		
NICE		
B.D.		
N.F.		
NeuroVL		
UMBRAE		
V.D.		
NeuroBind		
FLORA		



F Additional Qualitative Results

To achieve cross-modal concept-level zero-shot retrieval, we realigned the THINGS-MEG dataset split to match THINGS-EEG2. Tab.6 demonstrates the impact of this realignment and trial averaging on retrieval performance. Notably, due to the realignment, some concepts in both training and test sets inherit different numbers of trials from the original MEG dataset structure. To maintain consistency, we standardized to one trial per concept in the "unave." setting. However, for concepts that were originally in the MEG test set (which had 12 trials per concept), we also evaluated a "averaged" setting where these 12 trials were averaged. The performance improvement from FLORA-unimodal (unave.) to FLORA-unimodal demonstrates the benefit of this trial averaging, which effectively improves the signal-to-noise ratio for these concepts.

Table 6. **Impact of average retrieval performance across multiple conditions on the THINGS-MEG test set.** We report the retrieval performance of FLORA on the THINGS-MEG dataset under various evaluation settings.

Model	2-way Top-1	4-way Top-1	10-way Top-1	200-way Top-1	200-way Top-5
FLORA-unimodal (unave.)	72.38	54.62	36.88	5.75	17.87
FLORA-unimodal	**81.75**	**64.50**	**46.62**	**8.00**	**24.38**
FLORA-multimodal (unave.)	74.75	52.00	36.38	6.12	18.25
FLORA-multimodal	80.50	61.88	39.75	6.88	23.38

The original THINGS-MEG dataset presents this evaluation scenario in its data partitioning: the training set contains 1,654 concepts (12 images per concept, single repetition), while the test set comprises 200 concepts (1 image per concept, 12 repetitions per image) in a concept-level zero-shot setting. In this experiment, FLORA-unimodal was trained without the MoE-based universal projection and multimodal training, using an independent model for each subject's MEG recordings in an in-subject evaluation protocol. While our model has demonstrated superior performance in cross-subject zero-shot scenarios with richer data availability (as shown in Tab.1), it maintains competitive performance in this more constrained setting, as evidenced in Tab.7. The results validate that FLORA's core architecture provides robust performance even when trained on single-subject data with limited per-concept samples, without leveraging the benefits of multimodal training data and cross-subject information.

Table 7. **Retrieval performance on the original THINGS-MEG dataset.** We trained and evaluated various methods on the original THINGS-MEG training and test sets. Each method was independently trained and assessed for each subject, with results reported for different retrieval configurations: 2-way, 4-way, 10-way, as well as Top-1 and Top-5 accuracy for the 200-way retrieval task.The methods compared include NICE (Song et al., 2023), EEGNetV4 (Lawhern et al., 2018), B.D., MindEyeV2 (Scotti et al., 2024b) and ATM-S (Li et al., 2024b).

Model	2-way Top-1	4-way Top-1	10-way Top-1	200-way Top-1	200-way Top-5
NICE	89.29	76.94	60.19	17.12	40.15
EEGNetV4	85.34	71.78	52.85	12.76	32.38
B.D.	77.16	57.76	37.82	6.06	19.38
ATM-S	**90.25**	**79.50**	**63.66**	**17.84**	**44.73**
FLORA-unimodal	87.75	75.25	57.50	15.62	39.37

Our evaluation demonstrates that FLORA achieves strong performance in both in-subject and cross-subject settings on the THINGS-EEG2 dataset. When tested in the in-subject paradigm, where models are independently trained and tested on individual subjects' data as is shown in Tab.8, FLORA maintains competitive performance.

Notably, our cross-subject evaluation results shown in Tab.1 show better retrieval performance compared to the in-subject setting, validating our model's effectiveness in addressing

Table 8. **In-subject Retrieval on THINGS-EEG2.** We report the retrieval performance of different methods on the THINGS-EEG2 (in-subject) datasets. Each method is trained and evaluated in each subject independently, with results for different retrieval configurations: 2-way, 4-way, 10-way, and the Top-1 and Top-5 accuracy of 200-way. The methods compared include NICE (Song et al., 2023), EEGNetV4 (Lawhern et al., 2018), B.D. (Benchetrit et al., 2024), MindEyeV2 (Scotti et al., 2024b), UMBRAE (Xia et al., 2024a), ATM-S (Li et al., 2024b), CongnitionCapturer (Zhang et al., 2024), MindBridge (Wang et al., 2024b), WAVE (Wang et al., 2024d) and MB2C (Wei et al., 2024).

Model	THINGS-EEG2 (in-subject)				
	2-way Top-1	4-way Top-1	10-way Top-1	200-way Top-1	200-way Top-5
NICE	93.23	83.93	69.22	21.67	51.34
EEGNetV4	91.42	80.21	63.37	16.84	42.58
B.D.	89.03	56.77	56.77	13.29	35.50
CogCap	93.15	82.85	69.35	22.05	51.60
MB2C	78.40	62.25	43.75	8.85	25.20
Wave	88.65	73.70	54.70	13.00	34.50
MindBridge	89.10	75.35	57.45	15.00	39.15
ATM-S	**94.70**	**86.73**	**74.00**	**26.85**	**57.21**
MindEyeV2	92.50	82.80	66.10	23.80	50.25
FLORA-unimodal	93.75	85.70	69.65	21.85	51.65

cross-subject challenges. FLORA's robust performance can be attributed to its efficient encoder architecture across different data scales and the flexibility to incorporate or remove components like the MoE module. These results confirm FLORA's capability to deliver strong performance across various experimental paradigms while effectively leveraging larger-scale datasets when available.

Table 9. **In-subject Retrieval on THINGS-MEG.** We report the retrieval performance of different methods on the THINGS-MEG dataset in the in-subject setting. Each method is trained and evaluated in each subject independently, with results for different retrieval configurations: 2-way, 4-way, 10-way, and the Top-1 and Top-5 accuracy of 200-way. The methods compared include NICE (Song et al., 2023), EEGNetV4 (Lawhern et al., 2018), B.D., MindEyeV2 (Scotti et al., 2024b) and ATM-S (Li et al., 2024b).

Model	THINGS-MEG (in-subject)				
	2-way Top-1	4-way Top-1	10-way Top-1	200-way Top-1	200-way Top-5
NICE (unave.)	68.65	43.63	24.88	2.63	9.25
EEGNetV4 (unave.)	69.25	49.63	29.00	4.37	13.25
B.D. (unave.)	56.75	35.00	16.12	0.75	5.37
MindEyeV2 (unave.)	68.12	45.13	24.62	2.25	9.37
ATM-S (unave.)	**80.13**	**60.75**	**42.75**	**7.38**	**21.75**
FLORA-unimodal(unave.)	76.25	59.12	37.25	6.25	17.25
NICE	79.75	60.75	40.00	6.00	20.50
EEGNetV4	80.37	62.87	41.00	6.75	21.50
B.D.	64.25	43.32	21.50	2.00	8.50
MindEyeV2	76.25	56.88	37.62	6.00	18.75
ATM-S	**82.87**	**68.00**	46.13	8.12	**27.12**
FLORA-unimodal	81.75	66.00	**48.13**	**10.25**	26.25

The in-subject evaluation on the THINGS-MEG dataset demonstrates that FLORA-unimodal achieves state-of-the-art performance in the averaged condition while maintaining robust performance in the unaveraged setting (Tab.9). This superior performance, particularly with high-quality averaged MEG signals, validates FLORA's strong decoding capabilities even when operating on smaller-scale datasets.

This consistent performance can be attributed to two fundamental aspects of our architecture: first, the model's adaptable design that effectively accommodates various experimental paradigms, and second, its ability to efficiently utilize the enhanced signal-to-noise ratios in averaged MEG data without relying on the advantages of large-scale datasets. These findings further substantiate FLORA's effectiveness as a versatile neural decoding framework, demonstrating robust performance across both joint-subject and in-subject experimental paradigms.

Table 10. **In-subject Retrieval on THINGS-fMRI.** We report the retrieval performance of different methods on the THINGS-fMRI dataset. Each method is trained and evaluated in each subject independently, with results for different retrieval configurations: 2-way, 4-way, 10-way, and the Top-1 and Top-5 accuracy of 200-way. The methods compared include NICE (Song et al., 2023), EEGNetV4 (Lawhern et al., 2018), B.D. (Benchetrit et al., 2024), MindEyeV2 (Scotti et al., 2024b), UMBRAE (Xia et al., 2024a), ATM-S (Li et al., 2024b), CongnitionCapturer (Zhang et al., 2024), MindBridge (Wang et al., 2024b), WAVE (Wang et al., 2024d) and MB2C (Wei et al., 2024).

Model	THINGS-fMRI (in-subject)				
	2-way Top-1	4-way Top-1	10-way Top-1	100-way Top-1	100-way Top-5
NICE	90.00	82.67	62.67	25.00	56.00
EEGNetV4	89.00	76.00	56.00	20.33	50.67
B.D.	80.33	69.67	33.33	9.67	24.00
CogCap	93.67	81.33	64.00	27.33	64.00
MB2C	93.00	80.33	61.33	23.67	57.00
Neuro.V2L	86.67	70.67	50.67	15.00	39.67
Wave	91.00	79.67	65.67	29.00	61.00
ATM-S	**95.00**	81.00	66.67	25.67	60.33
UMBRAE	89.00	77.00	60.00	22.00	50.67
MindEyeV2	**95.00**	86.67	**76.67**	**32.67**	**69.00**
FLORA-unimodal	92.33	**87.00**	67.00	24.33	61.67

The in-subject evaluation on the THINGS-fMRI dataset demonstrates FLORA-unimodal's strong adaptability in decoding capabilities, even when operating on limited single-subject data. As shown in Tab.10, our model maintains competitive performance, validating its effectiveness across varying data scales and modalities. These results substantiate that FLORA, while primarily optimized for cross-subject scenarios with larger-scale datasets, retains robust decoding capabilities even when constrained to limited single-subject fMRI data. This finding further underscores the versatility of our architectural design across diverse experimental paradigms.

Examining in-subject retrieval performanceTab. 8, 9, and 10 demonstrate FLORA's consistent performance across EEG, MEG, and fMRI modalities in in-subject settings, showing

that our model, while optimized for cross-subject generalization, maintains competitive performance when trained on single-subject data. Comparing with Tab.1 , notably, while traditional approaches often suffer significant performance degradation in cross-subject scenarios, FLORA maintains or improves its performance when moving from in-subject to joint-subject settings, proposing a solution to joint-subject large neural models.

G Additional images results

We present a qualitative comparison of the highest-, moderate-, and lowest-fidelity generated images in Fig. 7. EEG data were randomly sampled from all subjects viewing 200 images, and corresponding EEG-derived embeddings were extracted to condition the image generation process. By computing the cosine similarity between the CLIP embeddings of the generated and original images, we identified six examples for each fidelity category. In the highest-fidelity group, the generated images exhibit strong semantic correspondence with the original images while effectively preserving low-level visual attributes. In the moderate-fidelity group, the generated images maintain semantic coherence with the original stimuli, yet exhibit partial alterations in low-level visual details. In the lowest-fidelity group, both semantic integrity and low-level visual structures are substantially distorted, leading to a pronounced divergence from the original images. See Fig. 8 and Fig. 9 for MEG and fMRI reconstruction results.

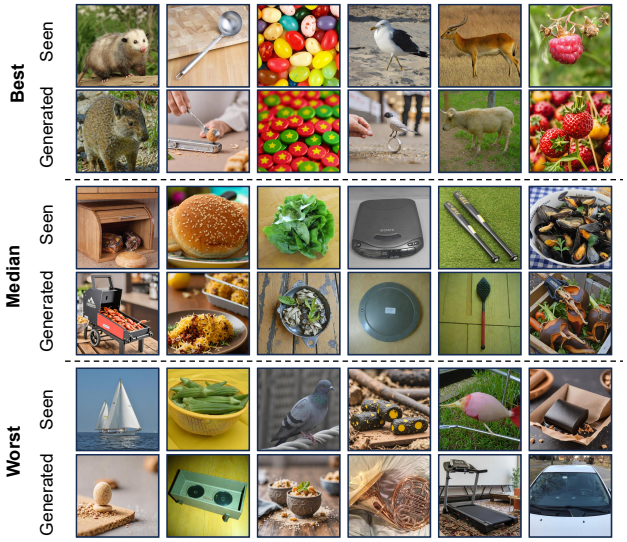


Figure 7. **Examples of cross-modality EEG-guided cross-subject visual reconstruction from FLORA**. From top to bottom, we exhibit the best, median, and worst 6 generated images, respectively. We show the images subjects seen and the generated images by our two-stage image generator.

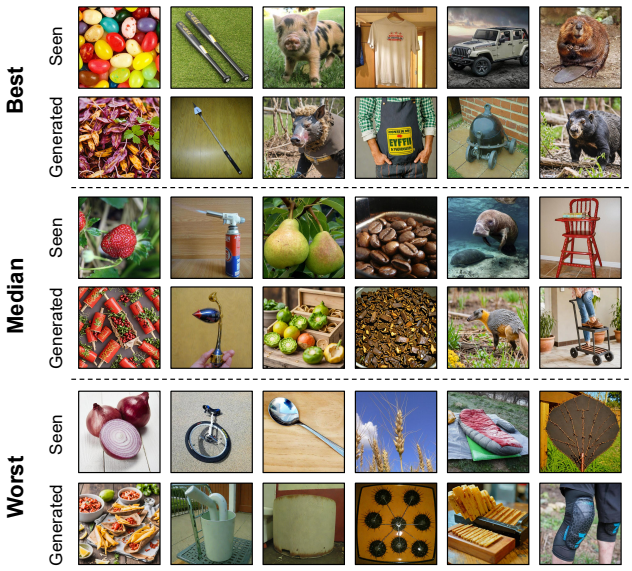


Figure 8. **Examples of cross-modality MEG-guided cross-subject visual reconstruction from FLORA**. From top to bottom, we exhibit the best, median, and worst 6 generated images, respectively. We show the images subjects seen and the generated images by our two-stage image generator.

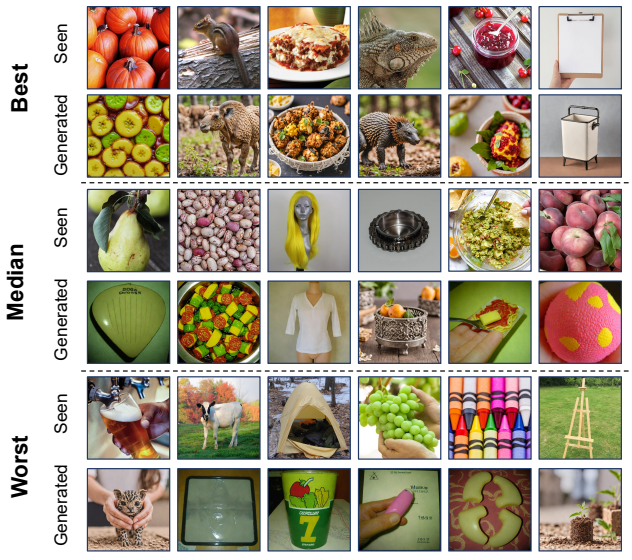


Figure 9. **Examples of cross-modality fMRI-guided cross-subject visual reconstruction from FLORA**. From top to bottom, we exhibit the best, median, and worst 6 generated images, respectively. We show the images subjects seen and the generated images by our two-stage image generator.