Cogformer: A unified multi-scale brain representation for visual decoding and reconstruction from fMRI

Xu Yin, John Q. Gan, Haixian Wang, *Senior Member, IEEE*

***Abstract*— With the rapid development of deep generative models (DGMs), the performance of decoding language and reconstructing images from Functional Magnetic Resonance Imaging (fMRI) has been improved. Nevertheless, the accurate representation of brain activity remains highly challenging, primarily due to the limited paired samples and the low signal-to-noise ratios of fMRI. To tackle these challenges, we introduce Cogformer, a unified multi-scale brain representation method. It is the first to learn brain representation from multi-scale fMRI activities via self-attention, and integrate a synchronized decoding and dynamic decoupling strategy for structural and semantic features through cross-attention. We conduct a systematic evaluation of Cogformer on the large-scale Natural Scenes Dataset (NSD) across a broad range of visual decoding tasks, including category classification, multi-label classification, image retrieval, image captioning, and image reconstruction. To the best of our knowledge, this represents the most extensive task coverage reported in related research. Cogformer achieves superior performance compared to a range of transformer-based baselines in category classification, multi-label classification, and image retrieval tasks. Moreover, in the more challenging tasks of image captioning and image reconstruction, Cogformer leverages a prior diffusion module to enhance the alignment with image semantics. This further improves the semantic consistency for caption generation and visual fidelity in image reconstruction. Across multiple evaluation metrics, Cogformer demonstrates competitive performance against existing state-of-the-art (SOTA) methods, highlighting its strong decoding capabilities and generalization potential.**

***Index Terms*— Functional Magnetic Resonance Imaging, Brain decoding, Transformer, Diffusion model**

I. Introduction

THE human brain is capable of rapidly constructing a comprehensive understanding of the visual world. This process involves not only the primary visual cortex, which encodes basic visual features such as edges, orientation, color, and depth, but also high-level brain regions responsible for recognizing and interpreting complex objects and scenes [1]. By recording brain activity using functional magnetic resonance imaging (fMRI) [2], [3], electroencephalography (EEG) [4], and electromyography (EMG) [5], researchers can progressively decode brain response patterns to visual stimuli, thereby uncovering the fundamental mechanisms of human visual cognition. fMRI offers high spatial resolution and has become widely used for decoding visual information. Visual neural decoding, which maps brain signals to visual perception, supports tasks like image recognition, semantic decoding, image captioning, and image reconstruction. Advances in computer vision (CV) and natural language processing (NLP) have further enabled multi-task decoding models [1], [4], [6], [7]. Regardless of modality or task, the challenge lies in learning accurate brain representations. Furthermore, the interpretability of brain representations is equally crucial for gaining neuroscientific insights. For example, Xia et al. [8] proposed DREAM, which employs carefully designed reverse pathways to emulate the hierarchical and parallel nature of human visual perception. These customized pathways are specifically tailored to decode semantics, color, and depth cues from fMRI data, thereby reflecting the forward pathways from visual stimuli to recorded brain activity. Similarly, Wei et al. [9] introduced MoRE-Brain, which adopts a hierarchical Mixture-of-Experts (MoE) architecture. In this framework, different experts are responsible for processing fMRI signals originating from functionally related voxel groups, thereby modeling specialized brain networks.

However, current brain representation methods face two major limitations. First, many approaches focus only on a single scale region of interest (ROI) [1], [8], [10], overlooking the multi-scale and hierarchical structure of the visual cortex. Although DREAM and MoRE-Brain have made valuable contributions to enhancing model interpretability, this limitation remains. DREAM emphasizes large-scale modeling of visual pathways, yet its interpretability at finer-grained levels is still limited. In contrast, MoRE-Brain adopts voxel-group experts as its core units, which provides advantages in local modeling but falls short of systematically uncovering the mechanisms of the visual cortex across middle-scale and large-scale brain regions. Cognitive neuroscience research has revealed that dorsal and ventral pathways in the brain regulate the flow of information within the visual cortex [11]. These two pathways reflect the hierarchical nature of visual information processing, highlighting the interconnectedness and functional

This work was supported by the National Natural Science Foundation of China under Grant 62176054.*(Corresponding author: Haixian Wang.)*

Xu Yin and Haixian Wang are with the Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Science & Medical Engineering, Southeast University, Nanjing 211189, Jiangsu, China (e-mail: yin_xu@seu.edu.cn; hxwang@seu.edu.cn).

John Q. Gan is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK (e-mail: jqgan@essex.ac.uk).

specialization of different brain regions within the visual cortex. However, cross-scale cooperation is also a fundamental mechanism in visual information processing. Studies have shown that cooperation among small-scale brain regions within the primary visual cortex is crucial for edge detection and contrast enhancement [12], while interactions between large-scale parietal cortex (dorsal pathway) and temporal cortex (ventral pathway) play a vital role in visual working memory and complex scene understanding [13]. Additionally, cross-scale cooperation exhibits distinct multi-scale features, including large-scale networks (e.g., the default mode network) and local connections, which together facilitate the processing of complex visual stimuli [14]. Recent functional connectivity analyses (such as watching movies or listening to stories) further emphasize the im-portance of multi-scale interactions between brain regions for the effective encoding and decoding of visual information [15]. The above studies all highlight the biological interpretability and necessity of effectively modeling the visual information transfer mechanisms across multi-scale brain regions.

The second limitation is the separation of structural and semantic representations, and existing models attempt to independently decode these two types of information to improve the performance of downstream visual tasks [4], [10], [16]. DREAM employs two independent modules to extract semantic and structural information from brain responses, whereas MoRE-Brain separately leverages a Variational Autoencoder (VAE) and Contrastive Language–Image Pretraining (CLIP) to obtain structural and semantic representations, respectively. In fact, there is mutual coupling between them, and treating them in isolation will reduce modeling efficiency and effectiveness [3]. For example, when the brain processes an image of a bird, semantic information helps the brain out-line the overall shape and appearance of the bird, while structural information aids in inferring the object's category based on its contours and details. Therefore, the brain can not only infer the missing contours of the bird based on semantic information but also recognize that it is an image of a bird based on the contour information.

To address these issues, we propose a fMRI-based visual decoding and reconstruction framework, as outlined in Figure 1. During training, a novel fMRI encoder is trained to align brain representation with high-level semantic and low-level structural features extracted by pre-trained CLIP [17] and VGG [18] image encoder, respectively. This fMRI encoder is trained in a task-agnostic manner, while downstream components such as diffusion-based image captioning and reconstruction are task-specific and only applied during inference. The primary contributions of this work are summarized as follows:

1) We propose a unified and task-agnostic brain representation learning framework, Cogformer, which learns a shared fMRI representation independent of downstream tasks, while enabling flexible integration with task-specific decoding modules for five visual decoding tasks, making it the most comprehensive framework to date.
2) We design ROI-wise multi-scale self-attention encoder and cognitive injection cross-attention encoder to unified model structural and semantic features from fMRI data, improving performance and interpretability.
3) A general pipeline is designed in which a prior diffusion is used to further align brain representations with semantic features, and the generated captions serve as conditional inputs for image reconstruction.

II. Related Work

A. Semantic Decoding and Visual Stimulus Reconstruction

Decoding visual semantics from brain activity evoked by visual stimuli has become a key focus in neuroscience and brain-computer interface (BCI) research. Early studies targeted single-category classification [19], while recent efforts have shifted toward multi-label decoding [20] and image captioning [21], enabled by advances in generative language models. In parallel, visual reconstruction has evolved from low-level shape recovery [2] to high-fidelity scene generation [10], [16] using GANs, VAEs, and diffusion models. These approaches reveal how different brain regions contribute to perceiving and reconstructing complex visual stimuli, supporting potential applications in BCIs and visual restoration.

B. Transformer

The Transformer architecture, first introduced by Vaswani et al. [22], has shown superior performance in capturing long-range dependencies compared to RNNs and CNNs. Its success in NLP led to applications in CV, such as the Vision Transformer (ViT), which learns spatial features by processing image patches with self-attention [23]. Transformers have also been applied to biological signal processing. For example, Wang et al. [24] introduced Medformer to capture multi-scale long-range dependencies in EEG and ECG signals, enhancing the model's ability to predict health conditions and pathological states. Jiang et al. [25] proposed LaBraM, a general EEG representation framework that enables rapid adaptation to various downstream EEG tasks.

III. Materials and Methods

A. Experimental Data and Preprocessing

a) NSD Dataset: The Natural Scenes Dataset (NSD) [26] comprises fMRI recordings from 8 subjects while they viewed images of natural scenes. All fMRI data in the NSD were collected at 7T using a whole-brain, 1.8-mm, 1.6-s, gradient-echo, echo-planar image (EPI) pulse sequence. Only 4 subjects (1, 2, 5, and 7) completed all 40 scanning sessions in the NSD, which enables the construction of a complete and balanced dataset for each subject. Although the number of subjects is limited, this setting follows common practice in high-resolution 7T fMRI studies where the acquisition burden per subject is substantial [3], [6], [7], [8], [10]. During the training phase, each subject was presented with 8859 unique images, each displayed for 3 seconds, resulting in a total of 24980 fMRI trials (with a maximum of 3 repetitions per image). During the test phase, 982 unique shared images were presented, generating 2770 fMRI trials. We utilized a general linear model (GLM) to estimate single-trial Z-score

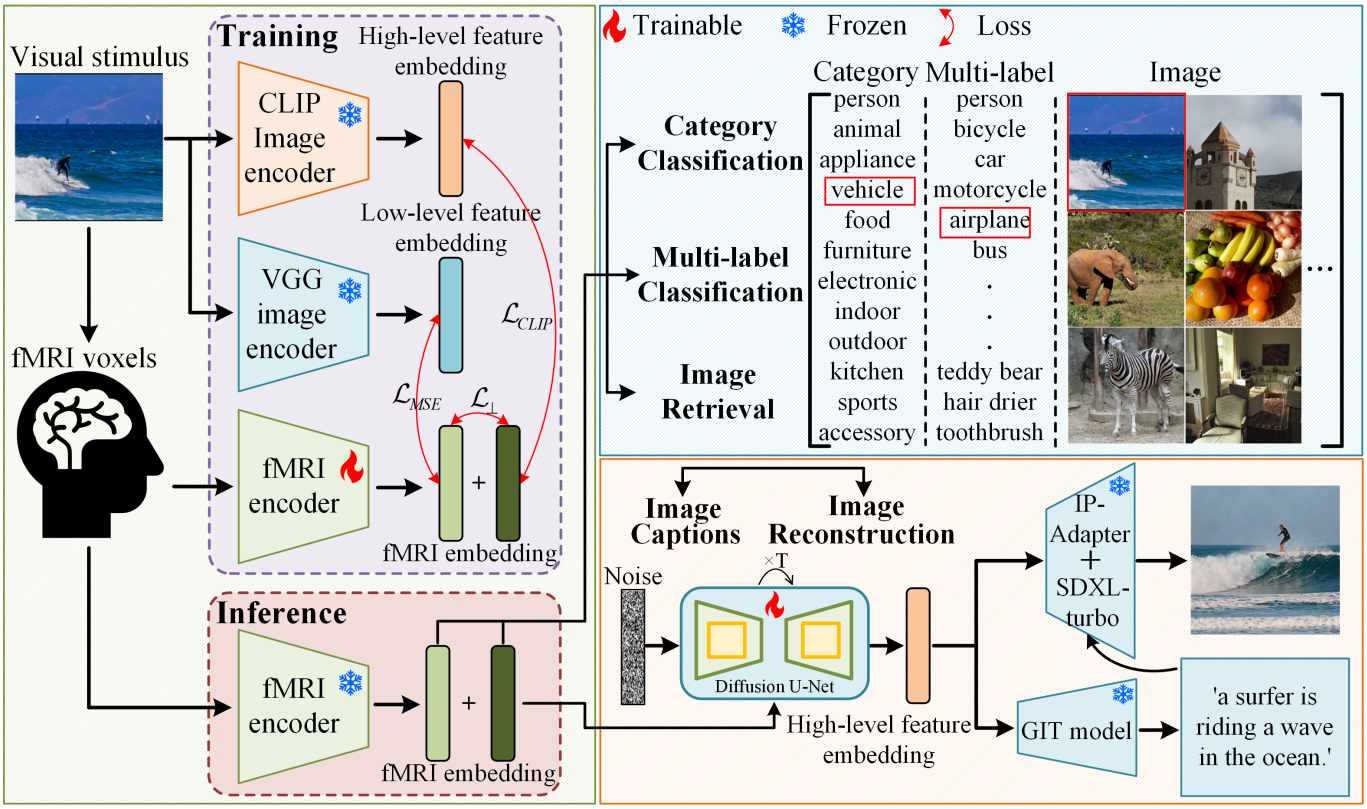

Fig. 1. fMRI-based visual decoding and reconstruction framework. In the training phase, an fMRI encoder is trained to align brain representation with high-level semantic and low-level structural features of images. In the inference phase, five tasks are carried out simultaneously: category classification, multi-label classification, image retrieval, image captioning, and image reconstruction.

beta weights and averaged multiple fMRI trails on repeated images to improve signal-to-noise ratio (For further details, see [26]).

The NSD dataset provides delineations of multiple regions of interest (ROIs) within the visual cortex. In this study, we selected 14 small-scale ROIs, each associated with well-established functional roles. Specifically, the Primary Visual Cortex dorsal part (V1d) and Primary Visual Cortex ventral part (V1v), together with the Secondary Visual Cortex dorsal part (V2d) and Secondary Visual Cortex dorsal part (V2v), process low-level visual features such as edges, orientation, and spatial frequency, while the Third Visual Cortex dorsal part (V3d) and Third Visual Cortex ventral part (V3v) further integrates these signals and contribute to form and motion representation. The human Visual Area 4 (hV4) is strongly linked to color and shape perception. High-level, category-selective regions include the Occipital Face Area (OFA) and Fusiform Face Area (FFA), which are specialized for face processing, the Occipital Word Form Area (OWFA) and Visual Word Form Area (VWFA), which are selective for word and orthographic representations, the Occipital Place Area (OPA), which is implicated in scene layout and spatial navigation, and the Extrastriate Body Area (EBA) and Fusiform Body Area (FBA), which are responsive to body parts and whole-body perception. The voxel counts of these ROIs are detailed in Table I. To capture hierarchical organization, we further grouped these ROIs into five middle-scale functional regions. The pre-visualrois correspond to early visual areas, including V1 through V3 and hV4, which support low-level feature encoding. In addition, four category-selective functional clusters were identified: floc-faces, which encompass OFA and FFA for face processing; floc-words, which include OWFA and VWFA for word processing; floc-places, which correspond to OPA for scene perception; and floc-bodies, which encompass EBA and FBA for body representation. This organization is consistent with well-established findings in cognitive neuroscience. At an even broader scale, we divided the visual cortex into two large-scale streams, following the dual-stream hypothesis. The dorsal stream, which consists of V1d, V2d, V3d, OPA, and EBA, is commonly described as the "where/how" pathway and is specialized for spatial localization, motion, and action-related processing. The ventral stream, which includes V1v, V2v, V3v, hV4, OFA, FFA, OWFA, VWFA, and FBA, is referred to as the "what" pathway and supports object recognition and semantic categorization. This multi-level grouping provides a biologically grounded framework for investigating visual cortical representations across scales.

The stimulus images were sourced from the COCO dataset [27], with the super-category providing 12 primary classes (e.g., person, animal, etc.). The multi-label annotations were derived from the 80 object names in COCO (e.g., person, bicycle, car, etc.), and the text descriptions were randomly selected from COCO's natural image captions. For each subject, we obtained 8859 training samples and 982 test samples. Each sample contains a natural image, a primary category label, multiple object labels, a textual description, and

corresponding fMRI responses, recorded across 14 small-scale ROIs, 5 middle-scale functional brain regions, and 2 large-scale visual streams.

b) Preprocessing: Brain activity exhibits trial-to-trial variability, even when responding to the same visual stimulus, posing challenges for stable decoding [21]. To mitigate this, we applied stability selection, identifying voxels with consistent activation patterns across repeated trials of the same image in the training set [28]. This was quantified using the Pearson correlation coefficient:

$$r = \frac{\sum(x - \bar{x})(x' - \bar{x}')}{\sqrt{\sum(x - \bar{x})^2} \cdot \sqrt{\sum(x' - \bar{x}')^2}} \tag{1}$$

where x and x' represent fMRI signals from two different trials of the same stimulus in the training set. Figure 2 provides a detailed visualization of this selection process. Since each image in training set involving 1 to 3 trials per subject, we standardized voxel selection by focusing only on images with three trials. The final number of stable voxels was constrained by the brain region with the fewest available voxels, resulting in 355, 441, 438, and 316 stable voxels for sub 1, 2, 5, and 7, respectively, as highlighted in bold in Table I.

B. Unified Multi-scale Brain Representation

To comprehensively capture the heterogeneity and hierarchical characteristics of visual cortical representations, we propose a unified multi-scale brain representation framework, termed Cogformer. As illustrated in Figure 3, this architecture is designed to jointly model intra-ROI and inter-ROI dependencies across multiple spatial scales, while dynamically disentangling low-level structural features and high-level semantic cues embedded in fMRI signals. The framework begins with a multi-scale patch embedding layer that partitions brain regions into small-scale, middle-scale, and large-scale groups, reflecting differences in functional granularity. These multi-scale brain embeddings are then processed through a multi-scale self-attention encoder, which captures both local interactions within each scale and global interactions across scales. To enhance model interpretability, we further introduce high-level and low-level tokens, which interact with the multi-scale brain representations via a cognitive injection cross-attention encoder, enabling explicit modeling of semantic and structural dimensions.

a) Multi-scale Patch Embedding: Previous fMRI processing methods leveraged architectures like RNNs [20], CNNs [29], and BiGRUs [1] to extract visual embeddings, but struggled to capture complex inter-regional interactions and long-range dependencies. To address this, we adopt a patch-based approach similar to Chen et al. [30], segmenting fMRI data into patches without relying on sequential recurrent processing. This enables Cogformer to better model global dependencies with improved efficiency. Additionally, we introduce a multi-scale patching mechanism to capture cross-region interactions at various spatial resolutions, inspired by multi-resolution strategies in Transformers [31] and medical time-series models [24]. Formally, given an fMRI signal $\boldsymbol{x}_{\text{in}} = (x_1, x_2, \ldots, x_T) \in \mathbb{R}^{V \times T}$ recorded while a subject views a natural image, where V represents the number of voxels from a single brain region and T denotes the number of multi-scale brain regions. We segment the input into $i-th$ scale patches $\boldsymbol{x}_p^{(i)} \in \mathbb{R}^{V \times (L_i \cdot C_i)}$ where L_i and C_i denote the length and number of patches at scale i. Each patch is then transformed into a lower-dimensional feature representation via an embedding function:

$$\boldsymbol{x}^{(i)} = \boldsymbol{x}_p^{(i)} \boldsymbol{W}^{(i)} \in \mathbb{R}^{D \times C_i} \tag{2}$$

where D represents the embedding dimension. This multi-scale patching strategy aims to comprehensively capture the hierarchical distribution of visual information in the brain, from fine-grained local features to a broader global perspective. In addition, we incorporate positional encoding [22] to explicitly introduce the spatial information of each brain region.

b) ROI-wise Multi-scale Self-attention Encoder: To capture the dependencies and interaction patterns between brain regions at different scales, we introduce a multi-scale self-attention mechanism. This module effectively integrates local and global visual representations by leveraging both intra-scale and inter-scale attention, allowing for a more comprehensive decoding of neural activity. For intra-scale self-attention, we employ a standard self-attention mechanism within each scale to capture local dependencies among patches at the same resolution. To enable efficient cross-scale information integration, we introduce a learnable connector embedding $c^{(i)}$ for each scale, serving as an information bridge that facilitates interactions across different hierarchical levels. For a given scale i, we concatenate the patch embedding $\boldsymbol{x}^{(i)} \in \mathbb{R}^{D \times C_i}$ with its corresponding connector embedding $\boldsymbol{c}^{(i)} \in \mathbb{R}^{D \times 1}$ to form an intermediate sequence representation:

$$\boldsymbol{u}^{(i)} = Concat(\boldsymbol{x}^{(i)}; \boldsymbol{c}^{(i)}) \tag{3}$$

where $Concat(\cdot)$ denotes concatenation along the sequence dimension. We then apply self-attention over this extended sequence:

$$\boldsymbol{x}^{(i)} \leftarrow Attention^{Intra}\big(\boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i)}, \boldsymbol{u}^{(i)}\big) \tag{4}$$

$$\boldsymbol{c}^{(i)} \leftarrow Attention^{Intra}\big(\boldsymbol{c}^{(i)}, \boldsymbol{u}^{(i)}, \boldsymbol{u}^{(i)}\big) \tag{5}$$

By assigning different attention weights to patches, this mechanism allows the model to dynamically capture long-range dependencies within each scale. Meanwhile, the connector embedding $\boldsymbol{c}^{(i)}$ is updated in the same way as the patch embedding $\boldsymbol{x}^{(i)}$, ensuring that it aggregates global information from within its respective scales. While intra-scale self-attention focuses on modeling fine-grained local interactions within each scale, inter-scale self-attention enables the integration of multi-level brain representations across different scales. We achieve this by first concatenating all connector embeddings across scales into a sequence:

$$\boldsymbol{C} = Concat\big(\boldsymbol{c}^{(1)}, \boldsymbol{c}^{(2)}, \ldots, \boldsymbol{c}^{(S)}\big) \tag{6}$$

where S represents the number of different scales. Each connector embedding $\boldsymbol{c}^{(i)}$ then attends to all other connectors using self-attention:

$$\boldsymbol{c}^{(i)} \leftarrow Attention^{Inter}\big(\boldsymbol{c}^{(i)}, \boldsymbol{C}, \boldsymbol{C}\big) \tag{7}$$

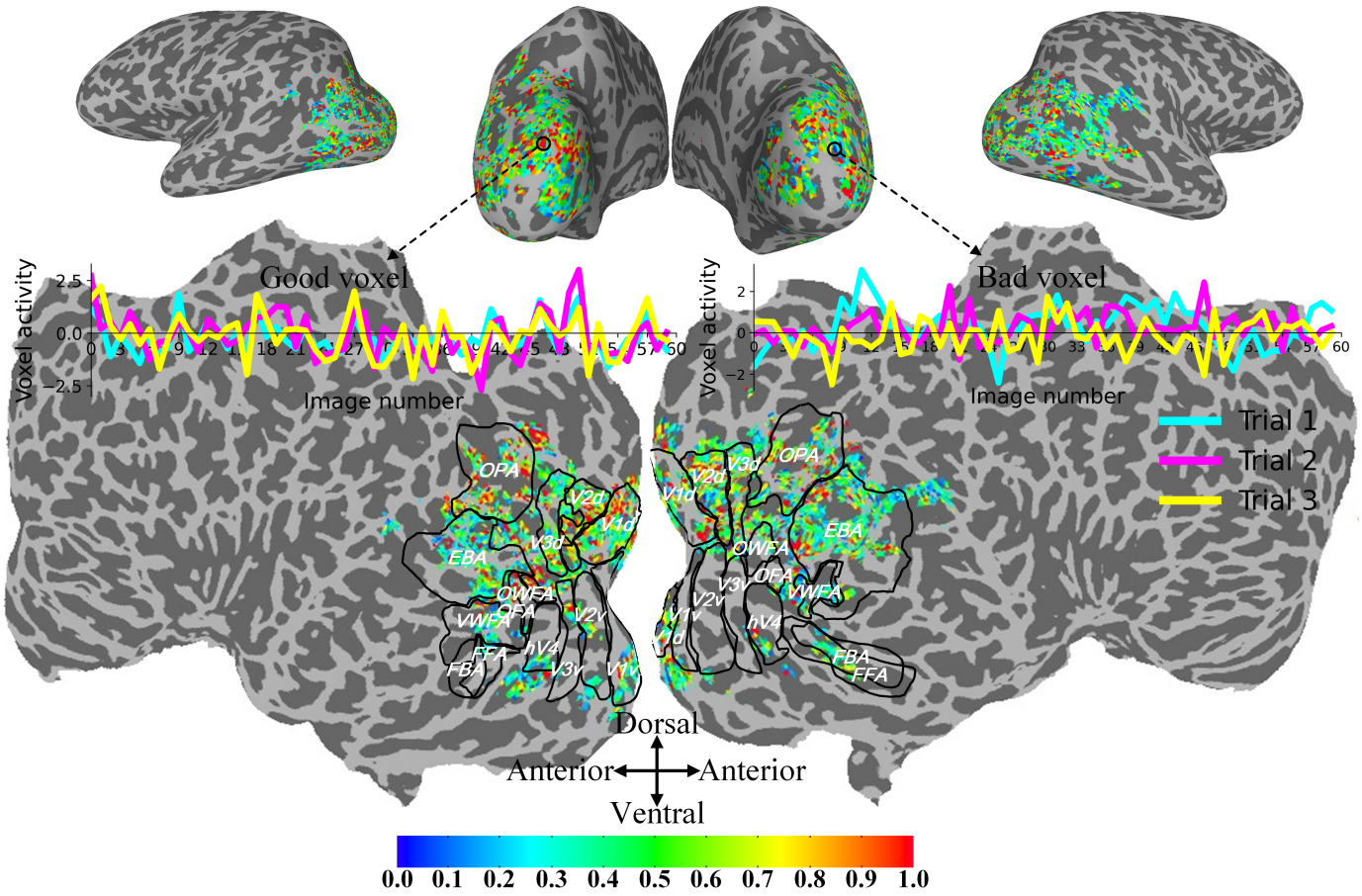

Fig. 2. The schematic illustration of voxel selection based on stability. The stability scores of all candidate voxels are mapped onto the cortical surface using a pycortex, with warmer colors indicating higher stability. We also compared the consistency across trials between selected good voxels and bad voxels.

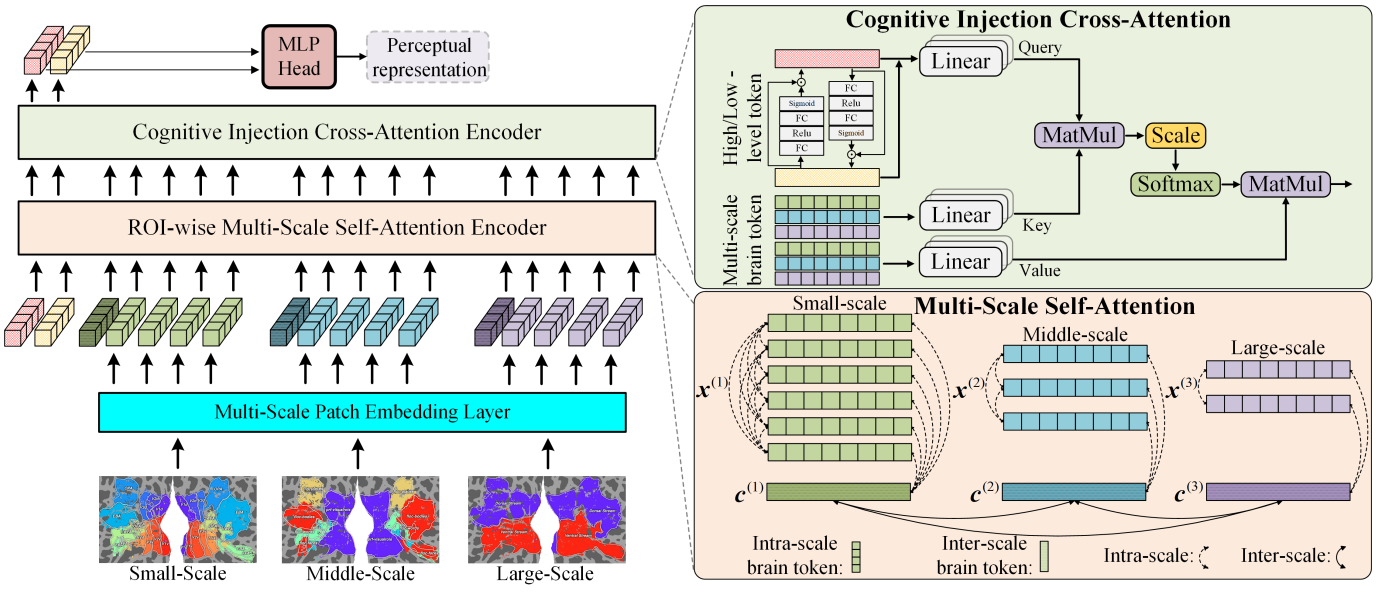

Fig. 3. The framework of unified multi-scale brain representation. Multi-scale patch embedding is employed to partition brain representations at different spatial scales; multi-scale self-attention is applied to capture interaction patterns both within and across scales; and cognitive injection cross-attention is introduced to dynamically decouple the representations of semantic and structural features.

TABLE I
THE DIMENSIONS OF RESPONSE ACTIVITIES (NUMBER OF VOXELS) OF SMALL-SCALE BRAIN REGIONS IN NSD.

Subjects	V1d	V1v	V2d	V2v	V3d	V3v	hV4	OFA	FFA	OWFA	VWFA	OPA	EBA	FBA
Sub 1	756	594	599	834	541	646	687	**355**	794	464	1083	1611	2971	826
Sub 2	558	544	460	615	531	566	483	**441**	869	519	821	1381	3439	1217
Sub 5	655	458	561	520	450	475	542	782	907	**438**	941	1332	4587	968
Sub 7	618	524	428	558	355	371	397	**316**	484	628	465	1083	3062	552

To further enhance representation capacity, we utilize multi-head self-attention, which allows the model to process information across different subspaces. For specific steps, please refer to [22]. After the attention sublayer, the encoder includes a fully connected feed-forward network. This component enhances the model's non-linearity and representational power by applying transformations to each token independently. The feed-forward network consists of two linear layers separated by a ReLU activation function.

c) Cognitive Injection Cross-Attention Encoder: In the human visual system, the processing of semantic and structural information is not entirely independent, but rather complementary and intertwined [32]. For example, when viewing an image of a bird, the brain can still infer the identity of the object as a bird even if parts such as the wings are occluded, by integrating shape and other structural cues. However, the interaction between semantic and structural features is not always fixed. In certain contexts, the brain may decouple these two types of information, particularly when specific aspects of an object or scene require focused attention [33]. When distinguishing between two bird species, the brain may prioritize structural features such as body size and shape, while decoupling the general semantic concept of "bird" from those features to enable more precise processing. These dynamic coupling and decoupling mechanisms allow the brain to flexibly adapt its representations in response to varying contextual and task-specific demands [34]. This flexibility in how the brain processes semantic and structural features inspires our proposed cognitive injection cross-attention encoder. In addition, we incorporate a learnable gating unit to dynamically modulate the degree of coupling between semantic and structural tokens. Specifically, given the high-level semantic token $\boldsymbol{z}_h$ and the low-level structural token $\boldsymbol{z}_l$, the gating mechanism produces their dynamically coupled representations $\tilde{\boldsymbol{z}}_h$ and $\tilde{\boldsymbol{z}}_l$:

$$\begin{aligned} \tilde{\boldsymbol{z}}_h &= \boldsymbol{z}_h + \sigma\left(MLP_l(\boldsymbol{z}_l)\right) \\ \tilde{\boldsymbol{z}}_l &= \boldsymbol{z}_l + \sigma\left(MLP_h(\boldsymbol{z}_h)\right) \end{aligned} \tag{8}$$

where σ is the sigmoid activation function, which ensures that the gating values lie within the range [0,1]. The Multilayer Perceptron (MLP) consists of two fully connected layers separated by a ReLU activation function. The resulting token $\tilde{\boldsymbol{z}} = [\tilde{\boldsymbol{z}}_h, \tilde{\boldsymbol{z}}_l]$ is then used as the query in the cross-attention, while the previously obtained brain tokens u serve as the key and value:

$$\tilde{\boldsymbol{z}} \leftarrow Attention^{Inter}(\tilde{\boldsymbol{z}}, \boldsymbol{u}, \boldsymbol{u}) \tag{9}$$

The cross-attention mechanism enables the token $\tilde{\boldsymbol{z}}$ to focus more effectively on brain activity patterns associated with visual stimuli, thereby retrieving both semantic and structural information. This mechanism can be viewed as a dynamic filtering process, where the semantic or structural token (serving as the query) selectively attends to brain region responses (serving as keys and values) based on its current state. Such a design allows the model to dynamically adjust its attention according to different visual inputs, akin to how the human brain flexibly modulates attention in varying contexts. By implementing a semantic and structural guided attention mechanism, the model is expected to achieve more accurate and interpretable representations of specific visual stimuli.

C. Training and Inference Process

a) fMRI Encoder Training: We choose CLIP and VGG as semantic and structural feature anchors, respectively, based on their complementary representational properties and extensive validation in prior work. CLIP is trained on large-scale image-text pairs and is known to encode high-level, category-discriminative and language-aligned semantic representations, making it particularly suitable for supervising semantic decoding tasks such as classification, retrieval, and captioning. In contrast, VGG features extracted from intermediate convolutional layers preserve fine-grained spatial, edge, and texture information, and have been widely used as perceptual or structural representations in image reconstruction and neural decoding studies. This combination allows us to anchor semantic and structural information to well-established and functionally distinct visual representations, facilitating effective disentanglement and alignment with fMRI signals. The fMRI encoder is trained by minimizing a total loss comprising three components: semantic alignment loss, structural alignment loss, and dynamic decoupling loss. The semantic alignment loss aligns the high-level semantic token $\tilde{\boldsymbol{z}}_h$ with the semantic representation $\boldsymbol{z}_{CLIP}$. Specifically, $\boldsymbol{z}_{CLIP}$ is obtained by feeding the image into the pre-trained CLIP image encoder (ViT-H/14, trained on LAION-2B). We directly use the 1024-dimensional image embedding produced by the projection layer, which corresponds to the projected global representation of the [CLS] token in the final image encoder output. A CLIP loss [17] (i.e., the InfoNCE loss [35] applied in both brain-to-image and image-to-brain directions) with exactly one positive example on batches of size N is calculated:

$$\begin{aligned} \mathcal{L}_{semantic} = -\frac{1}{N}\sum_{i=1}^{N}\Bigg[&\log \frac{\exp\left(s(\tilde{\boldsymbol{z}}_h^i, \boldsymbol{z}_{CLIP}^i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(s(\tilde{\boldsymbol{z}}_h^i, \boldsymbol{z}_{CLIP}^j)/\tau\right)} \\ &+ \log \frac{\exp\left(s(\tilde{\boldsymbol{z}}_h^i, \boldsymbol{z}_{CLIP}^i)/\tau\right)}{\sum_{k=1}^{N}\exp\left(s(\tilde{\boldsymbol{z}}_h^k, \boldsymbol{z}_{CLIP}^i)/\tau\right)}\Bigg] \end{aligned} \tag{10}$$

where $s(\cdot)$ denotes the cosine similarity, N and τ are the batch size and learnable temperature parameter, respectively. The structural alignment loss aims to align the low-level structural token $\tilde{z}_l$ captures fine-grained features by aligning it with structural representation z_{VGG}. Specifically, z_{VGG} is obtained by feeding the image into the pre-trained VGG16 feature extractor, followed by a global average pooling operation to produce a 512-dimensional image embedding. A mean squared error (MSE) on batch of size N is calculated:

$$\mathcal{L}_{structure} = \frac{1}{N}\sum_{i=1}^{N} \|\tilde{z}_l - z_{VGG}\|^2 \tag{11}$$

It is worth noting that we employ the CLIP loss (a contrastive learning objective) to optimize semantic representations, thereby enhancing their discriminability and cross-modal alignment. This objective plays a dominant role in training tasks such as category classification, multi-label classification, and image retrieval. In parallel, we use the MSE loss to constrain structural representations, ensuring the preservation of low-level edge, texture, and spatial information, which contributes to fine-grained local fidelity. This complementary design not only achieves a better balance between semantic and structural representations but also brings potential performance gains for the image reconstruction task. The dynamic decoupling loss introduces a gating mechanism that adaptively regulates the interaction strength between $\tilde{z}_h$ and $\tilde{z}_l$. Unlike the fixed orthogonality constraint proposed by Zhou et al. [3], this mechanism dynamically adjusts the degree of constraint based on the average gate value $\bar{g}$. When the interaction is strong, orthogonality is moderately relaxed to preserve feature complementarity; when the interaction is weak, orthogonality is reinforced to maintain representational independence. This mechanism effectively prevents gradient conflicts and avoids the rigidity of static constraints, thereby enhancing flexibility and robustness across multi-task settings and diverse types of visual stimuli:

$$\mathcal{L}_{orth} = (1-\bar{g})\left\|\tilde{z}_h^\top \tilde{z}_l\right\|^2 \tag{12}$$

Finally, we achieve accurate brain representation by minimizing the total loss:

$$\mathcal{L} = \mathcal{L}_{semantic} + \mathcal{L}_{structure} + \mathcal{L}_{orth} \tag{13}$$

Although no explicit scalar weighting is applied to each loss, the dynamic decoupling mechanism serves as an implicit, adaptive task weighting function, balancing the gradient contributions of different objectives during training.

b) Prior Diffusion Training: Inspired by Mind's Eye [36] and ATM [4], we train a diffusion model conditioned on the high-level semantic token $\tilde{z}_h$ derived from fMRI, aiming to further align its distribution with that of z_{CLIP}. This step is essential for image captioning and reconstruction, as contrastive learning only encourages the fMRI embeddings to match the vector direction of the corresponding CLIP image embeddings, which leads to disjoint embeddings [37]. Notably, this diffusion training stage is performed independently from the fMRI encoder training, and does not interfere with the multi-task optimization of semantic and structural objectives. During the forward process, we treat the z_{CLIP} as the initial state z_0 and progressively corrupt it into Gaussian noise via a fixed noise schedule:

$$q(z_t \mid z_{t-1}) = \mathcal{N}\left(z_t; \sqrt{1-\beta_t}\, z_{t-1}, \beta_t \boldsymbol{I}\right) \tag{14}$$

By recursively introducing noise, we obtain the feature representation z_t at any timestep t, which can be explicitly expressed in terms of the initial semantic feature z_0 and random noise $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$ as follows:

$$z_t = \sqrt{\bar{a}_t}\, z_0 + \sqrt{1-\bar{a}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, \boldsymbol{I}) \tag{15}$$

where $\bar{a}_t = \prod_{s=1}^{t}(1-\beta_s)$ controls cumulative noise scaling. At $t = T$, z_T approximates a standard Gaussian distribution. The reverse process learns to denoise z_T back to z_0, conditioned on the fMRI feature $\tilde{z}_h$. A U-Net predicts the distribution of z_{t-1} at each timestep:

$$p_\theta(z_{t-1} \mid z_t, \tilde{z}_h) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t, \tilde{z}_h), \sigma_t^2 \boldsymbol{I}) \tag{16}$$

where $\mu_\theta(\cdot)$ is predicted by a neural network conditioned on z_t, timestep t, and $\tilde{z}_h$; σ_t^2 may be fixed or learnable. We follow the standard DDPM objective, minimizing the MSE between the predicted noise and true noise:

$$L_{diffusion} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tilde{z}_h)\|_2^2\right] \tag{17}$$

c) Image Captioning and Reconstruction Inference: It is worth noting that the proposed fMRI encoder is a task-agnostic module trained independently of any specific downstream task, whereas task-specific modules are introduced only at the inference stage for image captioning and reconstruction. In the image captioning task, the 1×1024 high-level semantic feature $\tilde{z}_h$, derived from the output of the prior diffusion, is first fed into a pre-trained image projector to generate a 257×1024 latent image representation. This representation is then passed to the GIT model [38] to directly generate natural language captions. For the image reconstruction task, we employ SDXL-Turbo [39], a lightweight variant of Stable Diffusion XL, as the image generation backbone, in conjunction with the IP-Adapter [40] module for conditional guidance. The IP-Adapter enables the transformation of high-level semantic embeddings into prompt embeddings that are compatible with the vision-language diffusion model, thereby steering the generation process to better align with the semantics and structure of the original image. Compared to conventional multi-step diffusion models, SDXL-Turbo adopts a single-step sampling strategy, significantly improving inference efficiency while maintaining high-fidelity image synthesis. In addition, we incorporate a text-guided semantic pipeline into the image reconstruction process. The IP-Adapter takes the output captions from the image captioning task as textual prompts to further guide the semantic-level reconstruction of the image.

D. Implementation Details

The implementation of Cogformer follows a carefully structured design to balance computational efficiency and model expressiveness. The model architecture consists of 3 encoder layers, each with 256 hidden units ($d_{model} = 256$) and 8 attention heads ($h = 8$), enabling parallelized self-attention

and cross-attention mechanisms. To mitigate overfitting, a dropout regularization of 0.3 is uniformly applied across the network. For the training of the Cogformer, the batch size was set to 128 and the number of epochs was set to 300. We updated the parameters using the Adam optimizer with $lr = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$. Additionally, an early stopping mechanism is implemented, terminating training if the validation loss does not improve for 10 consecutive epochs, thereby preventing unnecessary computation and overfitting. In the prior diffusion module, the U-Net architecture adopts an encoder with layer dimensions set to [1024, 512, 256, 128, 64], with a decoder that mirrors this configuration in reverse. The timestep embedding dimension is set to 512. All hidden layers in the network use the Sigmoid Linear Unit (SiLU) as the activation function, defined as $SiLU(x) = x \cdot sigmoid(x)$. This function provides smoother gradients and stronger nonlinear representational capacity, leading to improved training stability in deep networks. In addition, the number of epochs was set to 150 and the dropout regularization was set to 0.1. We updated the parameters using the Adam optimizer with $lr = 0.0001$. To improve training stability and convergence, we utilize a cosine learning rate scheduler with 500 warm-up steps. All experiments were conducted on a high-performance computing workstation equipped with a 12th Gen Intel (R) Core (TM) i7-12700K CPU and one NVIDIA GeForce RTX 3090 GPU. All neural network pipelines were developed using the publicly available Pytorch framework, ensuring reproducibility and flexibility in model deployment.

IV. Results

A. Performance Evaluation Metrics

To evaluate the decoding performance across category classification, multi-label classification, and image retrieval, we first compute cosine similarities between the predicted high-level semantic token $\tilde{\boldsymbol{z}}_h$ and the semantic features $\boldsymbol{z}_{CLIP}$, as well as between the predicted low-level structural token $\tilde{\boldsymbol{z}}_l$ and the structural features $\boldsymbol{z}_{VGG}$. Based on these similarity scores, obtain similar retrieval images and calculate the following metrics: (1) Category Classification Accuracy: We report top-k accuracy by checking whether the ground-truth category appears among the categories of the top-k most similar retrieved images ($k = 1, 5$). (2) Multi-label Classification Accuracy: For each sample, we aggregate the label vectors of the top-k most similar retrieved images and compute the mean average precision (mAP) across all categories ($k = 1, 5, 10$). Here, for each category, we rank the predictions and progressively accumulate true positives (TP) and false positives (FP) to construct the precision–recall (PR) curve. We then compute the average precision (AP) from the curve, and finally take the mean across all categories to obtain the mAP. (3) Image Retrieval Accuracy: A retrieval is successful if the ground-truth image is included among the top-k most similar retrieved images ($k = 1, 5, 10$).

Unlike the evaluation metrics that rely on the cosine similarity between brain representations and image features, the performance of image captioning and image reconstruction tasks is assessed based on the predicted and ground truth text and images. For the image captioning task, we comprehensively evaluate the quality of generated captions using BLEU-n [41], METEOR [42], ROUGE-L [43], CIDEr [44], and SPICE [45]. For the image reconstruction task, features are extracted from both reconstructed and original images using pretrained models such as Inception [46], EffNet-B [47], SwAV [48], and CLIP [17], and their similarity is measured using metrics such as cosine similarity or Euclidean distance.

B. Category Classification, Multi-label Classification and Image Retrieval

To evaluate the performance of Cogformer on category classification, multi-label classification, and image retrieval, we compare it with four transformer-based brain representation baselines, each employing a distinct token embedding strategy for processing fMRI voxel responses. The input data $\boldsymbol{x}_{in} \in \mathbb{R}^{V \times T}$ undergoes different embedding processes in each method, as illustrated in Figure 4. To guarantee fairness in comparison, all methods are trained and evaluated on a common set of selected voxels derived from the same preprocessing pipeline.

(i) Transformer: The original Transformer [22] is primarily designed for sequence modeling. Here, it is extended to fMRI representation learning by treating individual voxels as tokens and establishing dependencies across the whole brain. This approach enables the capture of long-range interactions and mitigates the information decay issues encountered by traditional RNNs when modeling large-scale voxel data.

(ii) iTransformer: Unlike single-voxel modeling, the iTransformer [49] treats small-scale brain regions as units, using the entire voxel sequence within each region as a single token. This design focuses on intra-region dependencies and substantially reduces the number of tokens, thereby balancing representation effectiveness and computational efficiency.

(iii) PatchVAT (patch-based voxel area transformer): Inspired by the Patch-based Spatio-Temporal Structure (PatchSTS) model [50], this method further divides each small-scale brain region into multiple voxel patches to capture finer-grained local spatial features. By simultaneously modeling interactions within patches and across regions, PatchVAT effectively integrates local and global information, enhancing the model's ability for multi-level semantic decoding.

(iv) PatchBRT (patch-based brain regions transformer): This method is designed by us, which extends iTransformer by applying token embedding at the level of middle-scale brain regions, rather than individual voxels or small-scale patches, which enables the model to capture dependencies among middle-scale brain regions, offering insights into global interactions at a more aggregated level. Furthermore, PatchBRT generates fewer tokens than iTransformer, improving computational efficiency while retaining critical information about brain-wide dynamics.

(v) Cogformer: Our method introduces a hierarchical token framework that operates across multiple spatial scales. Rather than constraining token embeddings to a single level, Cogformer captures cross-scale dependencies among brain regions through a multi-scale self-attention mechanism. While

this design leads to higher computational costs compared to iTransformer and PatchBRT, the trade-off is justified by its ability to model the hierarchical nature of the human visual system. In addition, a cognitive injection cross-attention encoder is incorporated to further learn unified semantic and structural representations from brain tokens, aiming to enhance visual decoding performance. For clarity, we refer to the method without the cognitive injection cross-attention encoder as 'Cogformer-Base'.

Table II reports the chance levels and corresponding results for all baselines and our method in the test set. For the computation of the chance level in the primary category classification task, we directly calculate the overall prediction accuracy across all samples. To account for the issue of class imbalance, we further incorporate the distribution of test samples as weighting factors, thereby obtaining a more reasonable baseline level. Under the Top-k classification setting, the formalized equation is given as follows:

$$Chance_{micro}^{(k)} = \sum_{c=1}^{C} p(c) \cdot \left[1 - (1 - p(c))^k\right] \tag{18}$$

where C denotes the total number of categories ($C = 12$), and $p(c)$ represents the empirical distribution probability of class c in the test set. In the case of $k = 1$, the chance level reduces $\sum_{c=1}^{C} p(c)^2$. For the multi-label classification task, the chance level is computed by first calculating the average precision (AP) independently for each class, followed by taking the arithmetic mean across all classes. Since the final result is averaged over all categories, this metric is not affected by class imbalance during evaluation. The chance levels are as follows: category classification with 17.0% for Top-1 and 56.1% for Top-5; multi-label classification with 1.3% for Top-1 (1/80), 6.3% for Top-5 (5/80), and 12.5% for Top-10 (10/80); and image retrieval with 0.1% for Top-1 (1/982), 0.5% for Top-5 (5/982), and 1.0% for Top-10 (10/982). The experimental results indicate that in these relatively simple decoding tasks, all methods significantly outperform chance levels, and our method further surpasses the baselines. Moreover, the design of the cognitive injection cross-attention encoder provides additional improvements in visual decoding performance. To further compare the computational efficiency of different models, Table III reports the number of parameters (#Params), the floating-point operations (FLOPs), and the training time per epoch under the same hardware environment. Overall, although the proposed multi-scale brain representation introduces a moderate increase in parameter size, it does not lead to a corresponding rise in computational complexity or training cost. Instead, through a more principled architectural design and an efficient cross-scale feature modeling mechanism, our model achieves substantial performance gains while maintaining a controlled computational budget. This ability to obtain significant improvements in neural decoding accuracy with only a limited increase in computational cost highlights the practical value and clear advantages of the proposed method.

Figure 5 presents the category classification and multi-label classification performance. First, we input the brain signals of Sub 1 into the fMRI encoder to obtain the corresponding

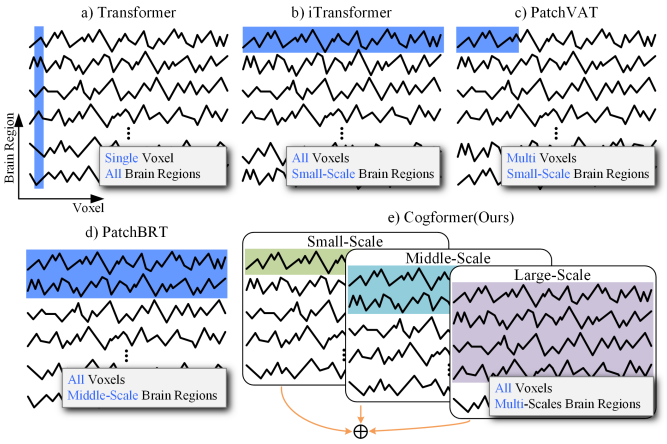

Fig. 4. The token embedding of different methods. Transformer utilizes a single voxel from all brain regions as a token; iTransformer employs all voxels from a small-scale brain region as a token; PatchVAT adopts multi voxels from a small-scale brain region as a token; PatchBRT uses all voxels from a middle-scale brain region as a token. To enhance multi-scale modeling of inter-region dependencies in the visual cortex, we propose a Cogformer that captures dynamic interactions across different brain regions.

perceptual representation $\tilde{z}$, whose two-dimensional t-SNE visualization is shown in Figure 5(a). The results demonstrate that the 12 major categories exhibit clear clustering structures with well-separated boundaries, confirming the strong semantic separability of the learned fMRI embeddings.To further investigate inter-class confusions, we plotted the confusion matrices of Sub 1's brain representations $\tilde{z}$ and the CLIP+VGG image features, as shown in Figure 5(b). The two modalities display highly consistent distribution patterns. Notably, semantic classes characterized by dynamic properties ("animate entities"), spatial scene attributes ("environment-related categories"), or fine-grained semantics such as "food" show similar prediction biases across modalities. Such structured confusions are not attributable to random noise but rather reflect the underlying semantic organization embedded in fMRI signals, which the model successfully captures and aligns. Figure 5 (c) presents the detailed classification accuracy for all 80 labels across the four subjects. Notably, labels such as "person" (50.63% of visual stimuli), "airplane" (2.87%), and "zebra" (4.32%) show higher average classification accuracy rates of 93.8%, 71.1%, and 62.6%, respectively. In contrast, labels with lower frequencies in the stimuli, such as hair drier (0.15%), toaster (0.21%), and scissors (0.21%), yield relatively lower decoding accuracies of 5.1%, 6.7%, and 7.2%, respectively. Nonetheless, the accuracy for all labels significantly exceeds the chance level of 1.3%. These results are highly consistent with the findings of Huang et al. [1], reinforcing the notion that the human brain encodes stable and distinct neural representations for different visual stimuli, which can be robustly captured by an effective decoding model.

C. Image Captioning

Table IV reports the average image captioning results across all subjects, comparing our method with several state-of-the-

TABLE II

COMPARISON OF THE ACCURACY (%) OF ALL METHODS IN CATEGORY CLASSIFICATION, MULTI-LABEL CLASSIFICATION, AND IMAGE RETRIEVAL. THE RESULTS ARE PRESENTED FROM BOTH TOP-1, TOP-5, AND TOP-10 PERSPECTIVES. THE HIGHEST CLASSIFICATION ACCURACY IN EACH COLUMN IS HIGHLIGHTED IN BOLD, WHILE THE SECOND HIGHEST IS UNDERLINED.

Methods	Category Classification Accuracy		Multi-label Classification mAP			Image Retrieval Accuracy		
	Top-1	Top-5	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
Chance Level	17.0	56.1	1.3	6.3	12.5	0.1	0.5	1.0
Transformer	63.9	78.6	24.0	43.0	56.4	23.0	57.0	67.9
iTransformer	64.6	79.2	24.6	43.7	58.0	24.4	57.5	69.4
PatchVAT	68.0	82.2	25.3	45.1	59.4	25.5	58.8	71.5
PatchBRT	68.6	82.5	26.8	45.8	60.8	<u>27.5</u>	59.9	73.2
Cogformer-Base	<u>70.3</u>	<u>84.4</u>	<u>28.4</u>	<u>47.3</u>	<u>61.4</u>	26.3	<u>60.4</u>	<u>74.9</u>
Cogformer	**71.3**	**86.1**	**30.3**	**48.6**	**63.4**	**29.3**	**62.3**	**76.1**

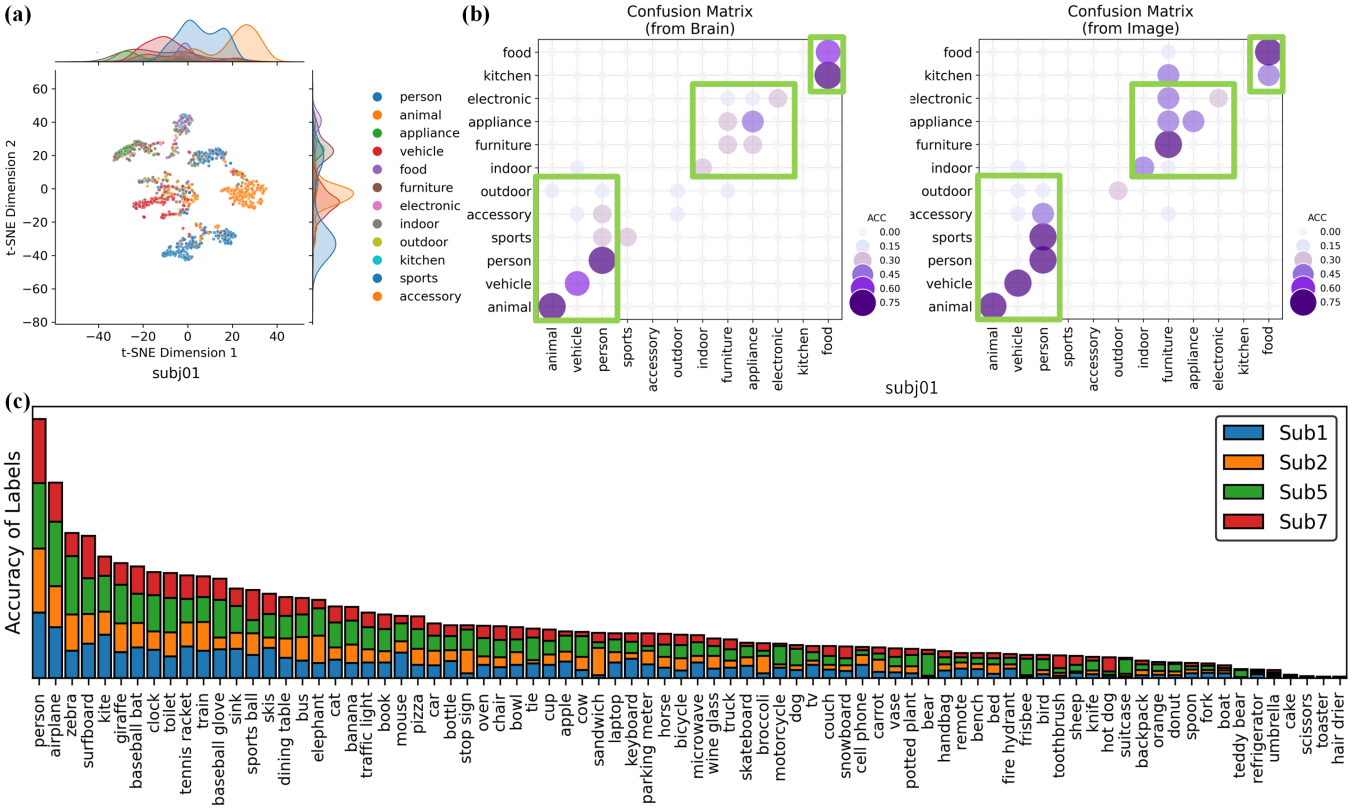

Fig. 5. Category classification and multi-label classification performance. (a) t-SNE visualization of perceptual representations learned from Sub 1 fMRI signals. (b) Confusion matrices of category from Sub 1 brain representation and image features. (c) Detailed classification accuracy for all 80 labels across the four subjects. Different colors represent different subjects and are sorted in descending order based on the total label accuracy for the four subjects.

TABLE III

COMPARISON OF MODEL PARAMETER COUNT (#PARAMS IN MILLIONS), COMPUTATIONAL COST (FLOPS IN MEGA/GIGA MACS), AND PER-EPOCH TRAINING TIME (SECONDS) UNDER THE SAME HARDWARE ENVIRONMENT.

Methods	#Params	FLOPs	Time per Epoch
Transformer	5.34M	1.11GMac	30.48s
iTransformer	5.25M	36.66MMac	4.74s
PatchVAT	3.98M	974.85MMac	22.83s
PatchBRT	5.25M	18.15MMac	4.71s
Cogformer	10.96M	82.57MMac	6.97s

art(SOTA) methods, including SDRecon [16], OneLLM [51], UniBrain [52], BrainCap [53], and UMBRAE [7]. As shown in the table, Cogformer outperforms all methods across all metrics except for CIDEr, where it is slightly outperformed by UMBRAE. SDRecon performs poorly in image captioning due to its limited vocabulary and the generation of redundant content. OneLLM improves caption quality by learning a unified encoder for multimodal text alignment. UniBrain and BrainCap enhance decoding performance by incorporating diffusion models. UMBRAE preserves richer semantic and spatial cues decoded from brain signals, enabling the generation of fluent, coherent, and informative sentences. In comparison,

our method not only integrates the strengths of the above models but also decodes rich semantic information from multi-scale brain regions and aligns it with visual semantics through prior diffusion, resulting in more accurate and descriptive captions.

Figure 6 compares predicted and ground-truth captions across subjects in the text description task. The model effectively captures core image semantics, generating contextually appropriate and meaningful descriptions. It shows strong performance in object recognition, scene understanding, and especially action depiction, as seen in accurate captions like "a man riding a surfboard on a wave in the ocean". The model also generalizes well in complex scenes, correctly incorporating elements like "tower" and "clock" into coherent sentences. However, limitations remain in capturing fine-grained details and distinguishing categories. Some captions oversimplify descriptions, such as mentioning only "a mirror" for a full bathroom scene or reducing "a couple of zebras" to "a zebra".

Fig. 6. Example of image captioning for all four subjects. Five images are randomly selected from the test set, displaying their ground-truth captions (in black) and predicted captions (in blue).

D. Image Reconstruction

Table V reports the average image reconstruction results across all subjects, comparing our method with several SOTA methods, including MindReader [54], SDRecon [16], Cortex2Image [55], BrainDiffuser [56], UniBrain [52], UMBRAE [7], MindEye [36], DREAM [8], and BrainCLIP [6]. As shown in the table, Cogformer achieves either the best or the second-best scores across all image reconstruction metrics. We find that the most competitive SOTA method compared to Cogformer is the MindEye, which adopts a fundamentally different strategy by using residual MLPs to directly learn representations from whole-brain voxels [36]. While this design theoretically preserves the finest-grained voxel-level spatial resolution, it also results in extremely high-dimensional representations and substantial computational costs, which are prone to overfitting under the limited sample size of the NSD dataset. In practice, MindEye employs separate high-level and low-level pipelines, where the semantic embedding has a dimensionality of 257×768 and the perceptual embedding is 4×64×64. Such high-dimensional features not only lead to a dramatic increase in model complexity but also reduce training stability. As shown in Table VI, the residual MLP in the low-level pathway contains 206M parameters, while the high-level pathway reaches 906M parameters. In contrast, Cogformer adopts a structurally organized ROI-based approach, in which voxel responses are first aggregated within functionally defined brain regions and then integrated across ROIs. This design represents a biologically motivated dimensionality reduction strategy, which also appears in [1], [8]. Although it may smooth out certain fine-grained voxel variations, it substantially enhances statistical robustness and interpretability while effectively mitigating overfitting risks. Cogformer uses a unified encoder architecture based on Transformer, which only contains 11M parameters, nearly two orders of magnitude fewer than MindEye, and relies on the diffusion prior only for spatial mapping. This significant reduction in the number of parameters highlights Cogformer's superior computational efficiency and scalability, and it also emphasizes an important trade-off: ROI-based modeling sacrifices some spatial resolution while providing greater stability, interpretability, and improved multi-task decoding performance. Furthermore, we compare Cogformer with several cross-subject neural decoding SOTA methods, including MindBridge [57], MindEye2 [58], and MindTuner [59]. These approaches typically construct a shared functional representation space or latent alignment mechanism across subjects, enabling brain responses from different individuals to be mapped into a unified latent space. This design substantially increases the effective amount of training data and enhances model generalization. As shown in Table VII, although cross-subject methods benefit from larger aggregated datasets and achieve advantages on certain metrics, Cogformer attains comparable reconstruction performance without relying on any cross-subject alignment, and even achieves the best semantic consistency as measured by CLIP. These results further demonstrate the efficiency and reliability of our proposed framework in single-subject decoding scenarios.

As illustrated in Figure 7, the image reconstruction results demonstrate the consistency and robustness of the Cogformer model across different subjects. Despite the inherent variability in brain activity across individuals, the reconstructed images preserve key semantic elements, scene layouts, and object configurations. For instance, images depicting street views,

TABLE IV
COMPARISON OF IMAGE CAPTIONING PERFORMANCE ACROSS ALL METHODS. THE RESULTS ARE AVERAGED ACROSS ALL SUBJECTS, WITH THE BEST SCORES HIGHLIGHTED IN BOLD AND THE SECOND-BEST SCORES UNDERLINED.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
SDRecon (2023)	0.352	0.164	0.072	0.033	0.098	0.246	0.135	0.049
OneLLM (2024)	0.470	0.270	0.155	0.095	0.136	0.351	0.230	0.063
UniBrain (2023)	—	—	—	—	0.169	0.222	—	—
BrainCap (2023)	0.548	0.351	0.220	0.140	0.162	0.404	0.389	0.088
UMBRAE (2024)	0.574	0.380	0.252	0.171	0.183	0.420	**0.517**	0.118
Cogformer	**0.623**	**0.446**	**0.310**	**0.192**	**0.225**	**0.487**	0.511	**0.145**

cyclists, or gatherings are reliably reconstructed with coherent visual semantics by all subjects. These results highlight Cogformer's ability to effectively decode and synthesize visual information from diverse neural patterns, showcasing its generalization capacity in image reconstruction.

Figure 8 presents representative examples of reconstructed images on sub 1 using our proposed pipeline and its ablation variants, clearly illustrating the impact of different configurations on image reconstruction quality. As both the prior diffusion module and text guidance are incrementally incorporated, the semantic performance of the reconstructions improves significantly. First, the reconstructed images show more accurate object shapes. Second, semantic consistency and the restoration of fine details are markedly enhanced, as seen in examples such as the reconstructions of the elephant and the person riding a horse. Third, the overall naturalness and realism of the images are improved. For example, in the reconstruction of airplanes, houses, and trains, images reconstructed without prior diffusion modules or textual guidance look more like cartoons or sketches than realistic images.

TABLE V
COMPARISON OF IMAGE RECONSTRUCTION PERFORMANCE ACROSS ALL METHODS. THE RESULTS ARE AVERAGED ACROSS ALL SUBJECTS, WITH THE BEST SCORES HIGHLIGHTED IN BOLD AND THE SECOND-BEST SCORES UNDERLINED.

Methods	Inception↑	CLIP↑	EffNet-B↓	SwAV↓
MindReader (2022)	0.782	—	—	—
SDRecon (2023)	0.760	0.770	—	—
Cortex2Image (2023)	—	—	0.862	0.465
BrainDiffuser (2023)	0.872	0.915	0.775	0.423
UniBrain(2023)	0.878	0.923	0.766	0.407
UMBRAE (2024)	0.917	0.935	0.700	0.393
MindEye (2023)	0.938	0.941	0.645	**0.367**
DREAM (2024)	0.934	0.941	0.645	0.418
BrainCLIP-VAE (2025)	0.842	0.946	—	—
Cogformer	**0.943**	**0.952**	**0.631**	0.377

E. Other Qualitative and Quantitative Results

a) Multi-scale attention connection patterns: To evaluate Cogformer's ability to extract semantic representations across brain regions at different scales, we analyzed its attention connectivity patterns under various stimulus categories. Prior research by Wang et al. [60] has shown that distinct brain regions exhibit significantly different neural representations for animate (living beings) and inanimate (objects) semantic

TABLE VI
PARAMETER COUNT COMPARISON BETWEEN COGFORMER AND MINDEYE. COGFORMER REQUIRES NEARLY TWO ORDERS OF MAGNITUDE FEWER PARAMETERS THAN MINDEYE FOR FMRI REPRESENTATION LEARNING AND RELIES SOLELY ON THE DIFFUSION PRIOR MODEL FOR SPATIAL MAPPING.

Method		Parameter Count
MindEye	Low Level	206M residual MLP + CNN decoder
	High Level	906M residual MLP + diffusion prior
Cogformer	Low+High Level	11M Transformer + diffusion prior

TABLE VII
COMPARISON OF IMAGE RECONSTRUCTION PERFORMANCE WITH CROSS-SUBJECT NEURAL DECODING SOTA METHODS. CROSS-SUBJECT METHODS LEVERAGE SHARED LATENT SPACES ACROSS INDIVIDUALS, WHILE COGFORMER OPERATES SOLELY IN A SINGLE-SUBJECT SETTING.

Methods	Inception↑	CLIP↑	EffNet-B↓	SwAV↓
MindBridge (2024)	0.924	0.947	0.712	0.418
MindEye2 (2024)	0.954	0.930	0.619	0.344
MindTuner (2025)	0.956	0.938	0.612	0.340
Cogformer	0.943	0.952	0.631	0.377

categories. As illustrated in Figure 9, we visualize the attention connections for "animate" categories (e.g., person, animal, sports) and "inanimate" categories (e.g., appliance, vehicle, furniture). The attention maps are derived from the final layer of the fMRI encoder, retaining only connections with weights ≥ 0.6. The results show that in the "animate" category, the small-scale region EBA and the middle-scale region floc-bodies exhibit the strongest connectivity, whereas in the "inanimate" category, the strongest connections are observed between OPA and floc-places. Cognitive neuroscience studies have shown that regions like EBA are primarily involved in encoding visual information related to the human body, selectively responding to body shapes and non-facial features [61], while regions like OPA are selectively activated by stimuli related to physical places and man-made environments, playing a key role in spatial perception and scene recognition [62]. These findings align with our observations, further supporting Cogformer's effectiveness in semantic decoding and its biological interpretability.

b) Semantic feature representation: To intuitively evaluate Cogformer's ability to model semantic representations in the brain, we constructed a voxel-wise neural encoding model.

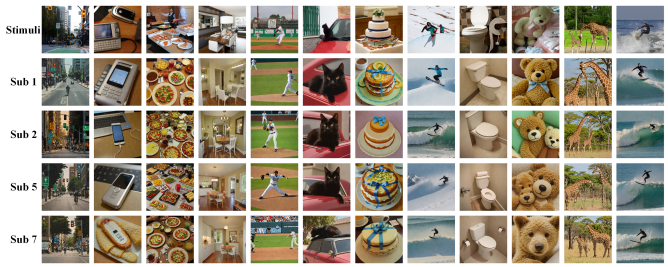

Fig. 7. Example of image reconstruction for all four subjects. Twelve images are randomly selected from the test set. The first row presents the original stimuli, while the second to fifth rows display the reconstruction results from the four subjects, respectively.

Fig. 8. Reconstructed images from our pipeline and its ablation models on sub 1. 'w/o PD w/o T' refers to the setting without the prior diffusion module and without text guidance. 'w/ PD w/o T' denotes the setting with the prior diffusion module but without text guidance. 'w/ PD w/ T' indicates the full version of our proposed pipeline, which incorporates both the prior diffusion module and text guidance.

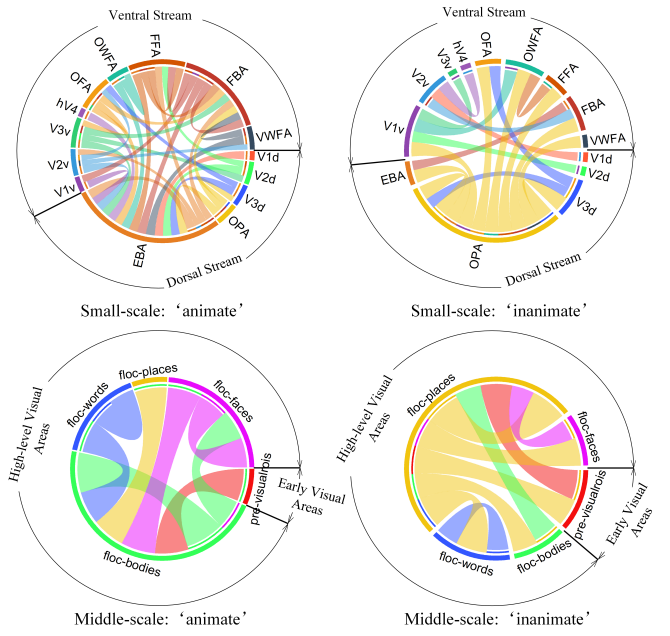

Fig. 9. The attention connectivity patterns between brain regions under different categories of visual stimuli, with explicit annotation of region groups. The left panel shows the connectivity pattern for the "animate" category, and the right panel shows the connectivity pattern for the "inanimate" category. Brain regions are grouped into dorsal and ventral streams for small-scale analysis (top row) and into early and high-level visual areas for middle-scale analysis (bottom row).

Specifically, high-level semantic token $\tilde{z}_h$, obtained from the prior diffusion module, were used to predict fMRI response of each voxel x_v through ridge regression. The model's performance was evaluated using the coefficient of determination R^2, which quantifies how effectively each voxel encodes visual semantic features. The encoding model was applied separately to stimuli from the "inanimate" and "animate" categories, and the difference ΔR^2 between the two conditions was computed and projected onto the 3D cortical surface using Pycortex [74], as shown in Figure 10. High-level visual areas such as EBA, FBA, and FFA showed negative ΔR^2 values, indicating a stronger response to animate stimuli. This result is consistent with their known involvement in biological form recognition. In contrast, regions such as OPA, OFA, OWFA, and VWFA exhibited positive ΔR^2 values, reflecting a preference for inanimate objects. Early visual areas such as V1 to V3 did not display a clear category preference.

c) Ablation study: To evaluate the effectiveness of the proposed unified brain representation and dynamic decoupling framework for semantic and structural features, we conducted a series of ablation studies. Specifically, Cogformer-L learns only low-level structural representations, Cogformer-H learns only high-level semantic representations, Cogformer-CONC directly concatenates semantic and structural representations without any decoupling, and Cogformer-ORTH enforces a strong orthogonality constraint between semantic and structural representations to achieve complete separation. The experimental results in Tables VIII and IX demonstrate consistent hierarchical differences in performance across these variants. Models relying on a single type of representation exhibited the lowest performance, with low-level structural features providing limited contributions, whereas high-level semantic features played a dominant role across all tasks. Incorporating structural information yielded additional gains, yet the fusion strategy substantially influenced overall effectiveness. Simple concatenation offered modest improvements but failed to effectively suppress cross-feature interference, thereby limiting its benefit. Orthogonal decoupling, by strictly separating the two types of features, mitigated interference to some extent, but its rigid constraints reduced the model's adaptability to varying coupling relationships under different stimulus conditions. In contrast, Cogformer, equipped with a dynamic decoupling mechanism, maintained the necessary independence between features while allowing flexible and task-relevant interactions. This design enabled the model to better capture multi-level associations between semantics and structure, achieving superior performance across all tasks.

In the ablation study of different scales of brain representations, we evaluated models using only small-scale ROIs, only middle-scale ROIs, several bi-scale combinations (small+middle, small+large, middle+large), and the full multi-scale integration implemented in Cogformer. Only large-scale ROIs were not tested individually because they consist of only two regions, which makes self-attention modeling impractical. As illustrated in Figure 11, Cogformer consistently achieved the best performance across all evaluation metrics, and paired two-tailed t-tests confirm the statistical significance of most differences (* indicates $p < 0.05$, **

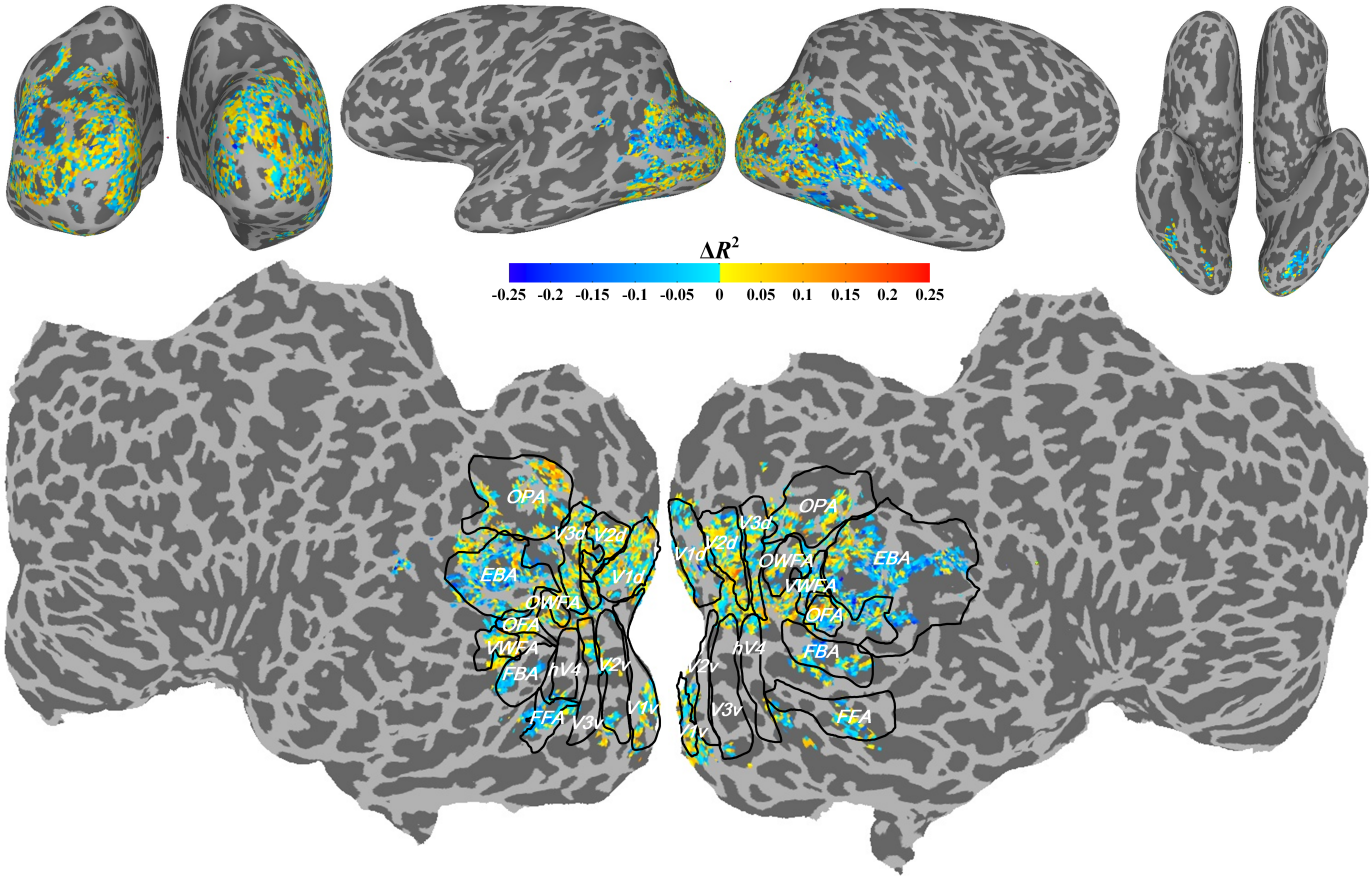

Fig. 10. Neural encoding performance of different brain regions for inanimate and animate visual stimuli. The projection of $\boldsymbol{\Delta R^2}$ onto the 3D cortical surface uses blue to indicate a preference for encoding animate stimuli and red to indicate a preference for inanimate stimuli. The darker the color, the greater the preference.

TABLE VIII

COMPARISON OF ABLATION EXPERIMENT RESULTS BY DIFFERENT BRAIN REPRESENTATION WAYS IN CATEGORY AND MULTI-LABEL CLASSIFICATION, IMAGE RETRIEVAL, AND IMAGE RECONSTRUCTION TASKS. THE RESULT IS THE AVERAGE OF ALL SUBJECTS, WITH THE BEST SCORES HIGHLIGHTED IN BOLD AND THE SECOND-BEST SCORES UNDERLINED.

Methods	Category and Multi-label Classification		Image Retrieval	Image Reconstruction			
	Accuracy	mAP	Accuracy	Inception↑	CLIP↑	EffNet-B↓	SwAV↓
Cogformer-L	44.0%	15.7%	17.2%	0.555	0.526	0.932	0.668
Cogformer-H	63.7%	24.6%	21.9%	0.764	0.849	0.738	0.591
Cogformer-CONC	68.5%	27.2%	25.6%	0.869	0.883	<u>0.670</u>	0.445
Cogformer-ORTH	<u>69.7%</u>	<u>28.1%</u>	<u>27.5%</u>	<u>0.877</u>	<u>0.893</u>	0.672	<u>0.422</u>
Cogformer	**71.3%**	**30.3%**	**29.3%**	**0.943**	**0.952**	**0.631**	**0.377**

TABLE IX

COMPARISON OF ABLATION EXPERIMENT RESULTS BY DIFFERENT BRAIN REPRESENTATION WAYS IN IMAGE CAPTIONING TASK. THE RESULT IS THE AVERAGE OF ALL SUBJECTS, WITH THE BEST SCORES HIGHLIGHTED IN BOLD AND THE SECOND-BEST SCORES UNDERLINED.

Methods	Image Captioning							
	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
Cogformer-L	0.301	0.217	0.109	0.044	0.080	0.209	0.216	0.023
Cogformer-H	0.519	0.322	0.215	0.113	0.146	0.355	0.370	0.092
Cogformer-CONC	<u>0.577</u>	0.398	<u>0.284</u>	0.155	0.174	0.388	0.409	0.111
Cogformer-ORTH	<u>0.577</u>	<u>0.401</u>	0.281	<u>0.157</u>	<u>0.188</u>	<u>0.413</u>	<u>0.435</u>	<u>0.126</u>
Cogformer	**0.623**	**0.446**	**0.310**	**0.192**	**0.225**	**0.487**	**0.511**	**0.145**

indicates $p < 0.005$, and *n.s.* indicates non-significance). Among single-scale settings, the middle-scale representation generally outperformed the small-scale one on metrics such as category classification accuracy (ACC_CateCls), mAP, image retrieval accuracy (ACC_ImgRet), and BLEU4, suggesting its advantage for semantic classification and captioning. The only exception was image reconstruction, where the CLIP similarity metric favored the small-scale representation. This likely reflects the small scale's finer parcellation and its better preservation of localized neural activity that benefits detail-sensitive reconstruction. Bi-scale combinations typically produce intermediate or improved performance relative to single-scale configurations, indicating complementary information across scales. In particular, the small+middle combination delivers the most consistent gains among the bi-scale variants, narrowing the gap to the full Cogformer on semantic metrics. This pattern suggests that combining fine-grained local signals with moderately aggregated regional signals enhances the model's ability to capture both local detail and broader semantic structure. The small+large combination shows relatively stronger benefits for reconstruction-related metrics, implying that adding coarse, global context from the large-scale ROIs can help integrate spatial layout or global perceptual cues important for reconstruction. The middle+large combination tends to provide modest improvements for semantic tasks over middle alone, but these gains are generally smaller than those obtained by small+middle. In summary, these observations demonstrate that the multi-scale fusion design of Cogformer is well justified, as different scales provide complementary signals and the learnable fusion leads to more robust support for multi-task decoding.

Previous studies have shown that the overall energy consumption level of the brain directly affects the functioning of key neural systems, including visual processing [63], [64]. Based on this finding, we designed an ablation experiment to investigate the trade-off between energy budget and task performance. Specifically, we introduced an energy penalty term into the total loss function to constrain attention energy consumption across different cortical scales, thereby assessing how reducing attention activation intensity influences multi-task visual decoding performance. The energy constraint term is defined as the mean attention connection weight across three scales and is balanced by a coefficient λ to regulate the trade-off between energy budget and task performance. To systematically analyze the model's behavior under different levels of energy constraint, λ was set in the range of 0 to 10, where $\lambda = 0$ represents the original model without any energy penalty. The model's performance was then evaluated across five task metrics, and the results of five repeated experiments are shown in Figure 12. We observe that as the energy constraint coefficient λ increases, corresponding to a tighter energy budget, the performance in category classification, image retrieval, and image captioning tasks remains relatively stable, while a more pronounced decline appears in the complex visual reconstruction task. This phenomenon may be attributed to the lower dependence of semantic decoding on energy consumption, whereas visual reconstruction requires stronger and more localized attentional activations to sustain detailed feature generation.

To evaluate the effectiveness of the dynamic decoupling mechanism under stimuli of different complexity levels, we selected multi-label samples containing annotated "person" to participate in the ablation experiment, and quantified the complexity of the stimuli through the number of labels. Figure 13(a) evaluates the impact of stimulus complexity on the degree of interaction between semantic and structural information. The results indicate that in complex stimuli containing a larger number of semantic labels (≥ 3 labels), the model tends to enhance the interaction between low-level and high-level representations to support the integration of global semantic information. In contrast, for simple stimuli with fewer labels (< 3 labels), the gating values were relatively lower, suggesting that the model places greater emphasis on maintaining the decoupling between structural and semantic representations to highlight local features and category-specific information. This trend demonstrates that the model can adaptively adjust its decoupling strategy according to the complexity of the input stimulus. This finding is also consistent with existing research in neuroscience and representation learning, which suggests that complex scenes demand stronger cross-level feature integration in neural representations, whereas simpler objects rely more on the separation of semantic and structural features [65]. To further analyze the effect of fixed gating under different stimulus complexities, we evaluated several fixed gate settings ($g = [0, 0.25, 0.5, 0.75, 1]$) and computed the corresponding image reconstruction CLIP scores, which were normalized to reflect their relative contributions. As shown in Fig. 13(b), fixed gating strategies exhibit clear performance differences across complexity levels. Lower g values, enforcing stronger semantic–structural decoupling, favor simple images, whereas higher g values, enabling stronger interactions, benefit semantically complex images. This behavior is consistent with the dynamic gating patterns observed in Fig. 13(a). Overall, the results indicate that the optimal interaction strength depends on stimulus complexity, and that dynamic gating, which adapts to the input, consistently outperforms any fixed gating strategy.

V. Conclusion and Future Work

In this work, we proposed Cogformer, a task-agnostic unified multi-scale brain representation framework based on fMRI. Cogformer learns brain representation from multi-scale fMRI activities via self-attention, and integrates synchronized decoding and dynamic decoupling strategy for structural and semantic features through cross-attention. Experimental results show that Cogformer consistently outperforms SOTA methods on five visual decoding tasks, including category classification, multi-label classification, image retrieval, image captioning, and image reconstruction, underscoring its powerful decoding ability and stable generalization across diverse tasks. Methodologically, the unified framework enables the simultaneous decoding of semantic and structural representations from fMRI signals, while dynamically decoupling the two during the learning process. This design allows the model to capture complementary information across multiple cognitive levels. In addition, the introduction of the prior diffusion further

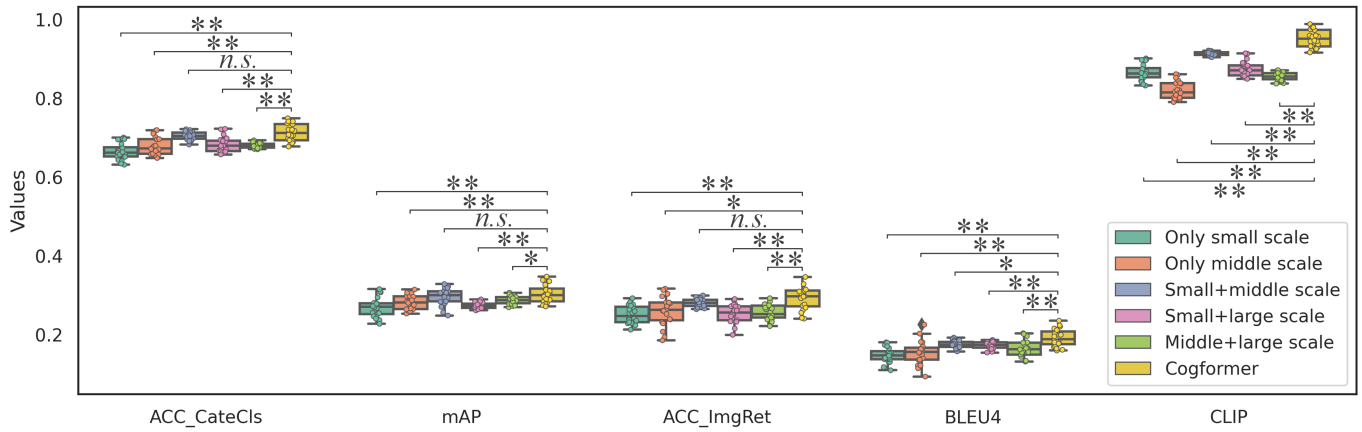

Fig. 11. Comparison of multi-task decoding performance across different scales of brain representation. Statistical significance was assessed using a paired two-tailed t-test, with * indicating significant differences ($p < 0.05$), ** indicating highly significant differences ($p < 0.005$), while *n.s.* indicates non-significant difference ($p \geq 0.05$).

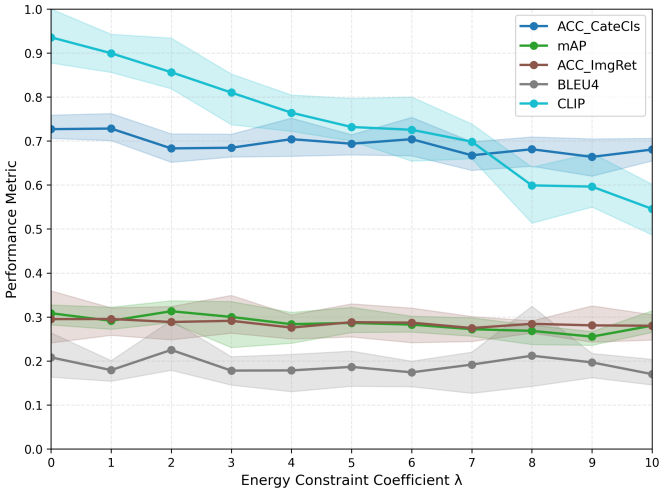

Fig. 12. Relationship between energy constraint Coefficient $\boldsymbol{\lambda}$ and task performance. Increasing $\boldsymbol{\lambda}$ imposes stronger energy regularization, resulting in stable performance (ACC_CateCls, mAP, ACC_ImgRet, BLEU4) across the semantic classification, image retrieval, and image captioning tasks, but a noticeable performance drop (CLIP) in the visual reconstruction task. Each point represents the average performance over five repeated experiments, and the shared envelope indicates the range between the maximum and minimum values.

aligns image semantics, which enhances the semantic consistency in image captioning and improves the visual fidelity in image reconstruction. Moreover, the visualization of attention connection patterns and semantic feature representation demonstrates the biological interpretability of the proposed method. Despite these significant advances, we acknowledge that the current study still has room for improvement in capturing finer low-level perceptual details. Future research can be extended in two directions: (1) incorporating more refined fMRI encoding mechanisms to enhance the representation of low-level visual features; and (2) integrating high-resolution reconstruction methods based on structural information to achieve collaborative modeling of low-level perception and high-level semantics, thereby advancing neural decoding to-

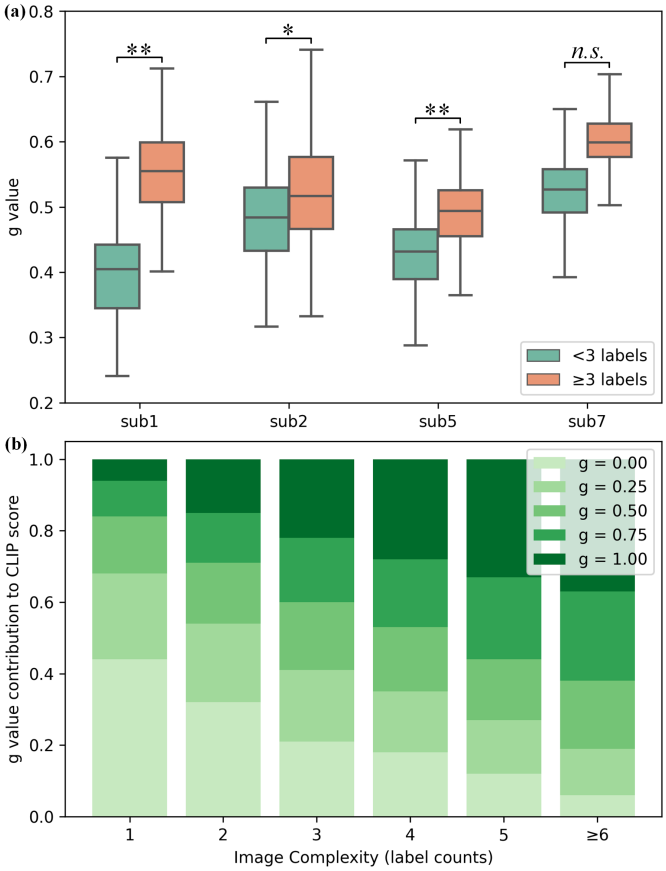

Fig. 13. Relationship between stimulus complexity and gating configurations. (a) Differences in gating values between simple (< 3 labels) and complex (≥ 3 labels) images, with statistical significance assessed using a paired two-tailed t-test (* $p < 0.05$, ** $p < 0.005$, *n.s.* $p \geq 0.05$). (b) Normalized contributions of fixed gating values to reconstruction performance CLIP scores across different image complexity levels.

ward higher accuracy and greater generalizability.

Although this study was conducted on 7T fMRI data (NSD), which offers a higher signal-to-noise ratio and finer spatial localization that facilitates multi-scale brain representation analysis, the proposed framework is not inherently dependent on high-field imaging. Previous studies [6], [30] have demonstrated that ROI-based brain representations and cross-modal alignment methods remain robust on 3T fMRI for semantic prediction and image reconstruction tasks, supporting the feasibility of visual decoding at lower field strengths. Given the widespread availability of 3T fMRI in both research and applied contexts, future work will extend evaluations to 3T datasets to assess the framework's generalizability and robustness across different field strengths. Moreover, substantial inter-individual variability in brain structure, functional organization, and cognitive style presents an additional challenge. Single-subject modeling often overlooks such neural heterogeneity, leading to limited cross-subject generalization. This lack of generalizability constrains the applicability of neural decoding in large-scale studies and real-world scenarios. Consequently, constructing unified brain representations across individuals has emerged as a key challenge and an important future direction in neural decoding research [58], [59], [66], [67].

Data Availability

Data is derived from an NSD (`https://cvnlab.slite.page/p/dC~rBTjqjb/How-to-get-the-data`). Code will be made public soon (`https://github.com/yinxu1996/visual-decoding`).

References

[1] W. Huang, et al. "From sight to insight: A multi-task approach with the visual language decoding model. Inform," Fusion, vol. 112, p. 102573, 2024.

[2] C. D. Du, C. Y. Du, L. J. Huang, H. G. He. "Reconstructing perceived images from human brain activities with Bayesian deep multiview learning," IEEE Trans. Neural Networks Learn. Syst., vol. 30, no. 8, pp. 2310-2323, 2019.

[3] Q. Zhou, C. Du, S. Wang, H. He. "CLIP-MUSED: CLIP-guided multi-subject visual neural information semantic decoding," In Proceedings of the International Conference on Learning Representations (ICLR), 2024.

[4] D. Li, C. Wei, S. Li, J. Zou, H. Qin, Q. Liu. "Visual decoding and reconstruction via EEG embeddings with guided diffusion," In Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), 2024.

[5] G. S. Mark, J. W. Michael, S. "Eelke. Decoding rich spatial information with high temporal resolution," Trends Cogn. Sci., vol. 19, no. 11, pp. 636-638, 2015.

[6] Y. Ma, Y. Liu, L. Chen, G. Zhu, B. Chen, N. Zheng. "BrainCLIP: Brain representation via CLIP for generic natural visual stimulus decoding," IEEE Trans. Med. Imaging., vol. 31, 2025.

[7] W. Xia, R.Charette, C. Öztireli, and J. Xue. "UMBRAE: Unified multimodal brain decoding," In Proceedings of the European Conference on Computer Vision (ECCV), 2024.

[8] W. Xia, R. Charette, C. Oztireli, J. H. Xue. "DREAM: Visual decoding from reversing human visual system," In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024.

[9] Y. Wei, Y. Zhang, X. Xiao, T. Wang, X. Wang, V. D. Calhoun. "MoRE-Brain: Routed mixture of experts for interpretable and generalizable cross-subject fMRI visual decoding," arXiv preprint arXiv:2505.15946, 2025.

[10] Y. Lu, C. Du, Q. Zhou, D. Wang, H. He. "MindDiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion," In Proceedings of the ACM International Conference on Multimedia (MM), 2023.

[11] M. Mishkin, L. G. Ungerleider, K. A. Macko. "Object vision and spatial vision: two cortical pathways," Trends Neurosci., vol. 6, pp. 414-417, 1983.

[12] X. Zhang, C. Lin, F. Li, Y. Cao, Y. Li. "LVP-net: A deep network of learning visual pathway for edge detection," Image Vision Comput., vol. 147, p. 105078, 2024.

[13] G. Guidali, C. Roncoroni, C. Papagno, N. "Bolognini. Cross-modal involvement of the primary somatosensory cortex in visual working memory: A repetitive TMS study," Neurobiol. Learn. Mem., vol. 175, p. 107325, 2020.

[14] B. R. Munn, et al. "Multiscale organization of neuronal activity unifies scale-dependent theories of brain function," Cell, vol. 187, no. 25, pp. 7303-7313, 2024.

[15] P. Fotiadis, L. Parkes, K. A. Davis, T. D. Satterthwaite, R. T. Shinohara, D. S. Bassett. "Structure-function coupling in macroscale human brain networks," Nat. Rev. Neurosci., vol. 25, pp. 688-704, 2024.

[16] Y. Takagi, S. Nishimoto. "High-resolution image reconstruction with latent diffusion models from human brain activity," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[17] A. Radford, et al. "Learning transferable visual models from natural language supervision," In Proceedings of the International Conference on Machine Learning (ICML), 2021.

[18] K. Simonyan, A. Zisserman. "Very deep convolutional networks for large-scale image recognition," In Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[19] T. Horikawa, Y. Kamitani. "Generic decoding of seen and imagined objects using hierarchical visual features," Nat. Commun., vol. 8, p. 15037, 2017.

[20] R. Li, et al. "Multi-semantic decoding of visual perception with graph neural networks," Int. J. Neural Syst., vol. 34, no. 4, p. 2450016, 2024.

[21] C. Du, K. Fu, J. Li, H. He. "Decoding visual neural representations by multimodal learning of brain-visual-linguistic features," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 9, pp. 10760-10777, 2023.

[22] A. Vaswani, et al. "Attention is all you need," In Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), 2017.

[23] A. Dosovitskiy, et al. "An image is worth 16x16 words: Transformers for image recognition at scale," In Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[24] Y. Wang, N. Huang, T. Li, Y. Yan, X. Zhang. "A multi-granularity patching transformer for medical time-series classification," In Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), 2024.

[25] W. Jiang, L. Zhao, B. Lu. "Large brain model for learning generic representations with tremendous EEG data in BCI," In Proceedings of the International Conference on Learning Representations (ICLR), 2024.

[26] E. J. Allen, et al. "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence," Nat. Neurosci., vol. 25, pp. 116-126, 2022.

[27] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al. Microsoft coco: Common objects in context. In Proceedings of European Conference on Computer Vision (ECCV), 2014.

[28] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, J. Schmidt, M. Shah. "Decoding brain representations by multimodal learning of neural activity and visual features," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 11, pp. 3833-3849, 2021.

[29] J. Zhang, et al. "A CNN-transformer hybrid approach for decoding visual neural activity into text," Comput. Meth. Prog. Bio., vol. 214, p. 106586, 2022.

[30] Z. Chen, J. Qing, T. Xiang, W. L. Yue, J. H. Zhou. "Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[31] Y. Zhang, L. Ma, S. Pal, Y. Zhang, M. Coates. "Multi-resolution time-series transformer for long-term forecasting," In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2024.

[32] K. Grill-Spector, K. S. Weiner. "The functional architecture of the ventral temporal cortex and its role in categorization," Nat. Rev. Neurosci., vol. 15, no. 8, pp. 536-548, 2014.

[33] S. Bracci, H. P. Beeck. “Dissociations and associations between shape and category representations in the two visual pathways,” J. Neurosci., vol. 36, no. 2, pp. 432-444, 2016.
[34] T. C. Kietzmann, C. J. Spoerer, L. K. A. Sörensen, R. M. Cichy, O. Hauk, N. Kriegeskorte. “Recurrence is required to capture the representational dynamics of the human visual system,” P. Natl. Acad. Sci. USA, vol. 116, no. 43, pp. 21854-21863, 2019.
[35] A. Oord, Y. Li, O. Vinyals. “Representation learning with contrastive predictive coding,” arXiv preprint arXiv:1807.03748, 2018.
[36] P. S. Scotti, et al. “Reconstructing the mind’s eye: fMRI-to-image with contrastive learning and diffusion priors,” In Proceedings of International Conference on Neural Information Processing Systems (NeurIPS), 2023.
[37] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen. “Hierarchical text-conditional image generation with CLIP latents,” arXiv preprint arXiv:2204.06125, 2022.
[38] J. Wang, et al. “Git: A generative image-to-text transformer for vision and language,” arXiv preprint arXiv:2205.14100, 2022.
[39] A. Sauer, D. Lorenz, A. Blattmann, R. Rombach. “Adversarial diffusion distillation,” In Proceedings of European Conference on Computer Vision (ECCV), 2024.
[40] H. Ye, J. Zhang, S. Liu, X. Han, W. Yang. “Ip-Adapter: Text compatible image prompt adapter for text-to-image diffusion models,” arXiv preprint arXiv:2308.06721, 2023.
[41] K. Papineni, S. Roukos, T. Ward, W. J. Zhu. “Bleu: a method for automatic evaluation of machine translation,” In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), 2002.
[42] S. Banerjee, A. Lavie. “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), 2005.
[43] C. Y. Lin. “Rouge: A package for automatic evaluation of summaries,” Text Summarization Branches Out, 2004.
[44] R. Vedantam, C. L. Zitnick, D. Parikh. “CIDEr: Consensus-based image description evaluation,” In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[45] P. Anderson, B. Fernando, M. Johnson, S. Gould. “Spice: Semantic propositional image caption evaluation,” In Proceedings of European Conference on Computer Vision (ECCV), 2016.
[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna. “Rethinking the Inception architecture for computer vision,” In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[47] M. Tan, Q. V. Le. “EfficientNet: Rethinking model scaling for convolutional neural networks,” In Proceedings of the International Conference on Machine Learning (ICML), 2019.
[48] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin. “Unsupervised learning of visual features by contrasting cluster assignments,” In Proceedings of International Conference on Neural Information Processing Systems (NeurIPS), 2020.
[49] Y. Liu, et al. “itransformer: Inverted transformers are effective for time series forecasting,” In Proceedings of International Conference on Learning Representations (ICLR), 2024.
[50] Y. Nie, N. H. Nguyen, P. Sinthong, J. Kalagnanam. “A time series is worth 64 words: Long-term forecasting with transformers,” In Proceedings of International Conference on Learning Representations (ICLR), 2023.
[51] J. Han, et al. “Onellm: One framework to align all modalities with language,” In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024.
[52] W. Mai, Z. Zhang. “Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity,” arXiv preprint arXiv:2308.07428, 2023.
[53] M. Ferrante, F. Ozcelik, T., Boccato, R. VanRullen, N. Toschi. “Brain captioning: Decoding human brain activity into images and text,” In Proceedings of International Conference on Neural Information Processing Systems (NeurIPS), 2023.
[54] S. Lin, T. C. Sprague, A. Singh. “Mind Reader: Reconstructing complex images from brain activities,” In Proceedings of International Conference on Neural Information Processing Systems (NeurIPS), 2022.
[55] Z. Gu, K. Jamison, A. Kuceyeski, M. Sabuncu. “Decoding natural image stimuli from fMRI data with a surface-based convolutional network,” In Proceedings of International Conference on Medical Imaging with Deep Learning (MIDL), 2023.
[56] F. Ozcelik, R. VanRullen. “Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion,” Sci. Rep., vol. 13, no. 1, p. 15666, 2023.
[57] S. Wang, S. Liu, Z. Tan, X. Wang. “MindBridge: A cross-subject brain decoding framework,” In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024.
[58] P. S. Scotti, et al. “MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data,” International Conference on Machine Learning (ICML), 2024.
[59] Z. Gong, et al. “MindTuner: Cross-subject visual decoding with visual fingerprint and semantic correction,” In Proceedings of the AAAI Conference on Artificial Intelligence, 2025.
[60] A. Y. Wang, K. Kay, T. Naselaris, J. T. Michael, W. Leila. “Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset,” Nat. Mach. Intell., vol. 5, pp. 1415-1426, 2023.
[61] M. G. Koningsbruggen, M. V. Peelen, P. E. Downing. “A causal role for the extrastriate body area in detecting people in real-world scenes,” J. Neurosci., vol. 33, no. 16, pp. 7003-7010, 2013.
[62] J. B. Julian, J. Ryan, R. H. Hamilton, R. A. Epstein. “The occipital place area is causally involved in representing environmental boundaries during navigation,” Curr. Biol., vol. 26, no. 8, pp. 1104-1109, 2016.
[63] D. Attwell, S. B. Laughlin. “An energy budget for signaling in the grey matter of the brain,” J. Cerebr Blood F. Met., vol. 21, no. 10, pp. 1133-1145, 2001.
[64] P. Lennie. “The cost of cortical computation,” Curr. Biol., vol. 13, no. 6, pp. 493-497, 2003.
[65] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, J. J. DiCarlo. “Performance-optimized hierarchical models predict neural responses in higher visual cortex,” Proc. Natl. Acad. Sci., vol. 111, no. 23, pp. 8619-8624, 2014.
[66] Y. Dai, et al. “MindAligner: Explicit brain functional alignment for cross-subject brain visual decoding with limited data,” In Proceedings of the International Conference on Machine Learning (ICML), 2025.
[67] Z. Wang, T. Pan, Z. Li, J. Wu, X. Li, J. Wang. “TROI: Cross-subject pre-training with sparse voxel selection for enhanced fMRI visual decoding,” In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025.