Aligning machines and minds: Neural encoding for high-level visual cortices based on image captioning task

Aligning Machines and Minds: Neural Encoding for High-Level Visual Cortices based on Image Captioning Task

Xu Yin[1], Jiuchuan Jiang[2], Sheng Ge[1], John Q. Gan[3], Haixian Wang[1, *]

[1] Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Science & Medical Engineering, Southeast University, Nanjing 211189, Jiangsu, People's Republic of China.
[2] School of Information Engineering, Nanjing University of Finance and Economics, Nanjing 210003, Jiangsu, People's Republic of China.
[3] School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United Kingdom.
* Author to whom any correspondence should be addressed.

E-mail: hxwang@seu.edu.cn

Keywords: functional magnetic resonance imaging, deep neural network, neural encoding, image caption task, attention

Abstract

Objective. Neural encoding of visual stimuli aims to predict brain responses in the visual cortex to different external inputs. Deep neural networks (DNNs) trained on relatively simple tasks such as image classification have been widely applied in neural encoding studies of early visual areas. However, due to the complex and abstract nature of semantic representations in high-level visual cortices, their encoding performance and interpretability remain limited. *Approach.* We propose a novel neural encoding model guided by the image captioning task (ICT). During image captioning, an attention module is employed to focus on key visual objects. In the neural encoding stage, a flexible receptive field (RF) module is designed to simulate voxel-level visual fields. To bridge the domain gap between these two processes, we introduce the Atten-RF module, which effectively aligns attention-guided visual representations with voxel-wise brain activity patterns. *Main results.* Experiments on the large-scale Natural Scenes Dataset (NSD) demonstrate that our method achieves superior average encoding performance across seven high-level visual cortices, with a mean squared error (MSE) of 0.765, Pearson correlation coefficient (PCC) of 0.443, and coefficient of determination (R^2) of 0.245. *Significance.* By leveraging the guidance and alignment provided by a complex vision-language task, our model enhances the prediction of voxel activity in high-level visual cortex, offering a new perspective on the neural encoding problem. Furthermore, various visualization techniques provide deeper insights into the neural mechanisms underlying visual information processing.

1. Introduction

Modeling the neural encoding of visual stimuli is a critical research paradigm aimed at uncovering how the human brain processes and interprets visual information [1, 2]. By examining the relationship between visual stimuli and brain activity, neural encoding models not only help uncover fundamental principles of cognitive neuroscience but also provide innovative insights and approaches for applications in medical diagnosis, human-computer interaction, and related fields. Compared to other neuroimaging techniques, such as electroencephalogram (EEG) [3, 4] and magnetoencephalography (MEG) [5, 6], functional magnetic resonance imaging (fMRI) offers richer and spatially precise information about brain activity [7-10], making it uniquely advantageous for building neural encoding models.

In recent years, with the rapid advancement of deep learning technologies, neural encoding models based on deep neural networks (DNNs) have steadily emerged as a research hotspot [11-13]. Previous studies have proposed several traditional encoding models based on handcrafted features. Ahonen et al. [14] proposed an efficient facial image representation by extracting local binary pattern (LBP) texture features. Nishimoto et al. [15] employed Gabor wavelets and motion energy features to predict voxel responses in the visual cortex. Huth et al. [16] used 1705 words to annotate video data and encoded high-level visual regions based on semantic annotation features. Compared to traditional methods, DNN-based encoding models leverage the hierarchical structure of neural networks to progressively extract and encode visual information, which has been validated from multiple perspectives for better simulating the hierarchical information processing patterns of the human visual cortex [17-19]. Such as St-Yves et al. [20] achieved neural encoding by extracting Gabor features or deep network features from visual stimuli, and then constructed a mapping between visual features and voxel activity. Wen et al. [12] utilized deep residual networks to extract features for visual encoding, achieving better prediction performance than AlexNet. Seeliger et al. [21] proposed a neural information flow (NIF) model, which represents neural information processing through a network of coupled tensors, each encoding the representation of the

sensory input contained in a region of interest (ROI). Wang et al. [22] introduced a framework based on a spiking convolutional neural network (SCNN) to achieve neural encoding in a more biologically plausible manner. Ma et al. [23] proposed a large-scale parameters framework with a sizable convolutional kernel for encoding visual fMRI activity. With the emergence of pre-trained large models, some studies have begun to rely on their powerful semantic representations, demonstrating remarkable advantages in semantic encoding tasks of high-level visual cortices [24]. However, such approaches overlook the structural feature processing mechanisms of low-level visual areas, which limits their ability to explain the hierarchical processing of visual information. This disconnection not only weakens model interpretability but also restricts its potential applications in neuroscience research.

Existing research clearly shows that DNNs, driven by simple tasks like image classification, can extract structural representations of images, such as edges and textures, through low-level networks [25, 26], thereby ensuring effective encoding of low-level visual cortices. However, these simple tasks only require identifying key targets in natural image scenes, which may cause the network to overlook smaller targets, backgrounds, and their relationships. This limitation makes it difficult for DNNs to capture the global semantics of visual scenes, resulting in restricted encoding performance for high-level visual cortices. In contrast, the Human Visual System (HVS) captures all key elements in a visual scene, not just the main target. It forms representations of the relationships between different targets, develops a global understanding of the scene, and subsequently focuses on specific regions or targets based on attention mechanisms and the requirements of corresponding visual tasks. Therefore, we propose introducing high-level visual tasks to guide DNNs in constructing image representations that emphasize advanced semantics. Specifically, the image captioning task (ICT) [27, 28] involves generating a sentence or paragraph to describe the image content, effectively "speaking from the picture".

In this paper, we propose a novel end-to-end neural network framework for modeling neural encoding in the high-level visual cortex. The rich visual representations provided by the ICT module ensure high voxel encoding performance. Conversely, the voxel encoding facilitates the prediction of brain activity and enhances the biological interpretability of ICT. To bridge the gap between ICT (machine) and voxel encoding process (mind) during model training, we further design an innovative attention-constrained receptive field (RF) module, termed "Atten-RF". The contributions of this paper are summarized as follows:

i) By incorporating the hierarchical representation of visual processing and integrating more complex computer vision tasks, our method significantly improves the neural encoding performance in high-level visual cortices.

ii) A novel attention-based RF module is introduced to bridge the domain gap between visual images and brain responses, which in turn enhances the biological interpretability of computer vision models.

iii) Various visualization techniques are employed to investigate the RF distributions and semantic encoding characteristics of high-level visual cortices, contributing to a deeper understanding of the brain's information processing mechanisms.

2. Related work

2.1. Visual neural encoding

Early studies on visual neural encoding have demonstrated that mapping deep neural network (DNN) image representations to cortical activity through sparse linear regression can effectively reveal the correspondence between artificial features and the hierarchical processing of the visual cortex. For example, Wang et al. [26] and Seeliger et al. [29] showed that the shallow and deep layers of standard convolutional neural networks (CNNs) exhibit stable correspondences with the primary and higher visual cortices, respectively. However, due to the abstract and complex nature of representations in higher visual areas, these models still face limitations in encoding performance. With the emergence of pre-trained large models, researchers have attempted to leverage more expressive features to improve neural encoding. For instance, Wang et al. [24] employed Contrastive Language-Image Pretraining (CLIP) features to enhance prediction performance in higher visual areas. Moreover, Transformer-based architectures such as the Vision Transformer (ViT) [30] have recently demonstrated strong representational power, capable of learning diverse feature types through different pre-training objectives, including cross-modal alignment as in CLIP and structural reconstruction as in Masked Autoencoders (MAE). This provides new opportunities to examine how pre-training strategies influence encoding performance. Nevertheless, it should be noted that methods such as CLIP and ViT are more oriented toward global modeling in terms of training objectives and feature organization, and their hierarchical features do not strictly correspond to the stepwise processing in the biological visual system, thereby limiting their interpretability in neuroscience research. In contrast, architectures with explicit hierarchical structures, such as AlexNet and Residual Networks (ResNet), offer clearer layer-wise gradients, which

are more conducive to analyzing the correspondence between artificial features and cortical hierarchy, and thus provide higher analytical value for the study of higher visual areas [11].

2.2. ICT and attention

ICT aims to generate semantically rich, human-readable natural language descriptions for a given image. As a key intersection of computer vision (CV) and natural language processing (NLP), this task has widespread applications in image understanding, intelligent search engines, and automated translation systems [31]. Traditional approaches, which often rely on template matching or handcrafted feature extraction, offer some interpretability but are significantly limited in terms of flexibility and semantic depth. With the rapid advancement of deep learning, neural network-based methods have become mainstream. State-of-the-art image captioning models typically employ an "encoder-decoder" architecture. CNNs are widely used to extract image features, while Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), are employed to generate natural language descriptions [32]. This end-to-end learning paradigm facilitates the automatic capture of semantic associations between images and text directly from data, resulting in significant improvements in the quality and diversity of generated descriptions.

Drawing inspiration from human attention patterns, recent studies have integrated attention mechanisms to improve models' ability to focus on specific regions of an image. For instance, Xu et al. [33] proposed a soft attention-based image captioning model that dynamically focuses on the image regions most relevant to the word being generated. This approach enabled a closer integration of visual and textual information. Anderson et al. [34] further developed this concept by introducing a Bottom-Up and Top-Down Attention model, which leverages object detectors to identify salient regions of an image. This method significantly enhanced the accuracy of generated captions and the ability to express fine-grained details. Yao et al. [35] incorporated graph structures to tackle the relational challenges between visual and textual modalities, using the attention mechanism to align critical information across both.

2.3. RF modeling

The RF models aim to characterize the spatial extent and functional properties of neural units in response to stimuli within their visual or auditory environment [36]. By studying RF models, we can gain valuable insights into how neurons process and encode external information during perceptual tasks such as vision and audition. Kay et al. [37] utilized a linear combination of Gabor basis functions with different spatial frequencies, orientations, and positions to predict voxel responses in the early visual areas, achieving promising encoding results. St-Yves et al. [20] proposed an encoding model based on the feature-weighted receptive field (fwRF), which assumes that each voxel has a fixed RF associated with the processing of features from a specific region of the image stimulus. By weighting and combining features from the fixed RF across all feature maps of the DNNs, this model outperformed previous methods and revealed that different voxels in the visual cortex have distinct RFs. In contrast to the rigid prior assumptions about RF described above, Wang et al. [25] proposed an RF estimation method with weaker prior constraints, enhancing the expressiveness and interpretability of the encoding model.

RF models describe the local sensitivity of each voxel to input stimuli by defining the range of information it receives. This approach enables a more intuitive understanding of how different voxels encode specific stimuli, thereby delving deeper into the functional organization structure of the brain. Xue et al. [38] introduced the fwRF framework to train high-performing encoding models for the ventral visual pathway. However, these voxel-wise encoding methods treat each voxel as an independent unit, overlooking the interactions and information exchange between different voxels. This simple encoding approach not only suffers from low efficiency but also fails to capture the information processing mechanisms of the entire brain. In recent years, ROI-wise encoding methods have emerged, where voxels within different functional regions of the brain are encoded simultaneously. Qiao et al. [39] designed an end-to-end ROI-wise convolution regression model, achieving more effective and efficient visual encoding compared to existing voxel-wise methods.

3. Method

To enhance the prediction of voxel activity in high-level visual cortices and to explore the encoding mechanisms of the HVS, we propose a novel end-to-end neural encoding network. The schematic diagram of the proposed framework is presented in Figure 1A. In the following, each component and the training process of the proposed model will be explained in detail.

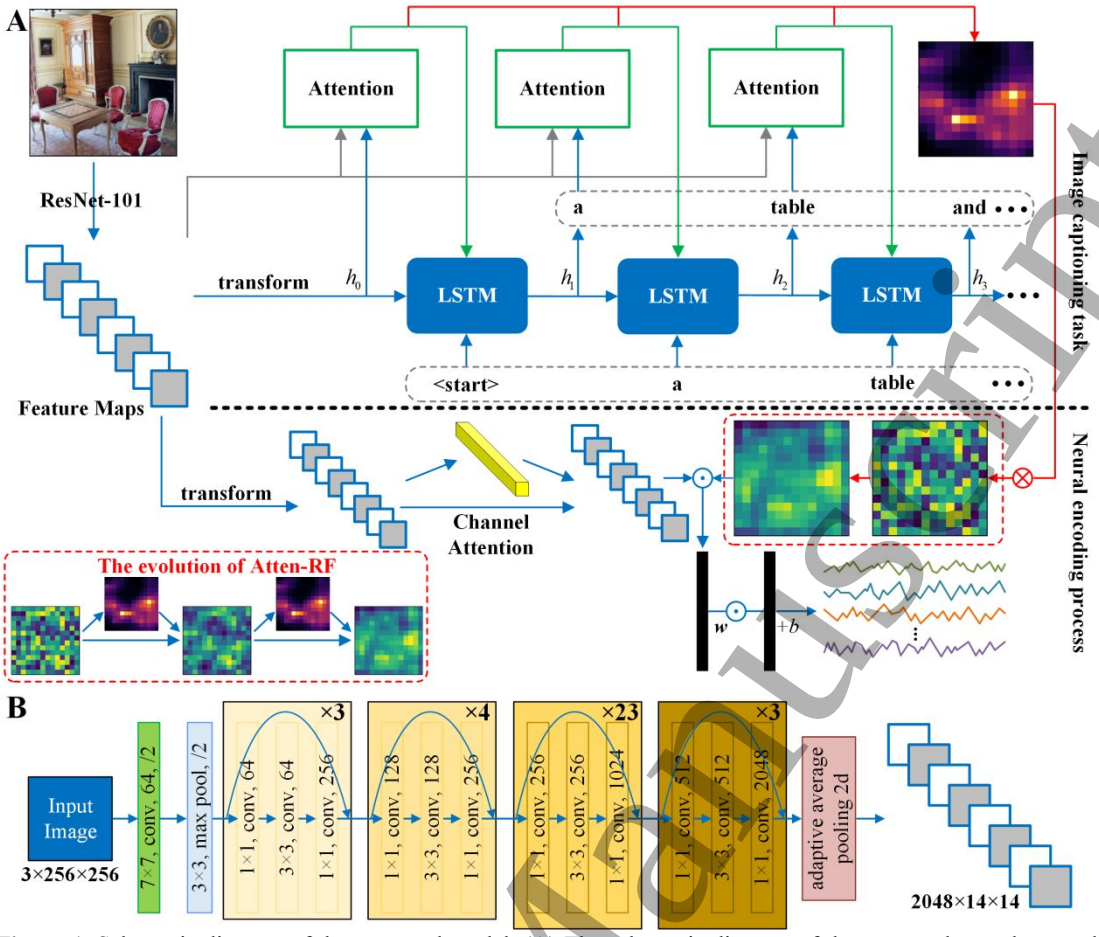

Figure 1. Schematic diagram of the proposed model. (A) The schematic diagram of the proposed neural network framework, comprising an ICT and a neural encoding progress. The "Atten-RF" module is designed to enhance consistency between the two during model training. (B) The feature map extraction module consists of ResNet101 without a classification head and an adaptive pooling layer.

3.1. Notations

Let N pairs of neural encoding data be represented as $\{(\boldsymbol{x}^n, \boldsymbol{t}^n, \boldsymbol{y}^n) \mid \boldsymbol{x}^n \in \mathbb{R}^P, \boldsymbol{t}^n \in \mathbb{R}^K, \boldsymbol{y}^n \in \mathbb{R}^V\}_{n=1}^N$, where $\boldsymbol{x}$ denotes the stimulus image, $\boldsymbol{t}$ represents the corresponding text, and $\boldsymbol{y}$ is the evoked fMRI activity. P, K, and V represent the dimensions of image pixels, text length, and the number of fMRI voxels, respectively. For a given image $\boldsymbol{x}^n$, the extracted feature maps are $\boldsymbol{FM}^n \in \mathbb{R}^{C \times S}$, where C represents the channel dimension and S denotes the spatial resolution. The matched spatial RFs of all voxels are defined as $\boldsymbol{rf} = [rf_1, \cdots, rf_V] \in \mathbb{R}^{V \times S}$. Ultimately, the predicted d-th word and v-th voxel's brain response are $\hat{t}_d^n$ and $\hat{y}_v^n$, respectively. Without loss of generality, we omit the sample annotation n in subsequent explanations and illustrate the process using a single sample.

3.2. Feature map extraction

Existing frameworks for neural encoding of visual stimuli universally rely on learning image representations, which transform visual stimuli into low-dimensional feature spaces to facilitate a deeper understanding and efficient processing of visual information. Over the years, numerous models have been built and pre-trained many models that are extraordinarily good at classifying an image into one of a thousand categories. Among these models, CNNs remain the mainstream framework, with their convolution, pooling, and related modules effectively capturing the spatial structural representations of image data. They progressively generate increasingly compact representations of the original image, with each subsequent layer producing more abstract features represented by a greater number of channels. We opted for the 101-layer Residual Network (ResNet101) [40], pre-trained on the ImageNet classification task, to extract feature maps and fine-tune them during model training, as this strategy typically achieves better performance than training a new model from scratch. The network is readily available in Pytorch, with its architecture illustrated in Figure 1B. As the last two layers are linear layers paired with a softmax activation for classification, we strip them away. In addition, to obtain a fine-grained spatial resolution,

which facilitates the generation of more detailed spatial attention maps and RF visualizations while maintaining computational efficiency, we further applied an adaptive average pooling layer to increase the resolution from 8×8 to 14×14. This operation essentially performs resolution upsampling in the spatial domain without introducing any new information beyond what is already present in the 8×8 features. Ultimately, our feature map extractor, based on ResNet101, generates feature maps $\boldsymbol{FM}$ with dimensions 14×14 and 2,048 channels.

3.3. Image captioning

Since the ICT involves generating a sequence, it requires an RNN. We employ an LSTM network, which effectively handles sequential data and long-term dependencies. Once generate the feature maps, we can simply average and transform them to initialize the hidden state h_0 and cell state c_0, enabling the LSTM to produce the word sequence. Each predicted word is used to generate the next one. Instead of the simple average, we want the LSTM to focus on different parts of the image at different timesteps in the sequence. For example, while generating the word "airplane" in the sentence "an airplane sits at the airport waiting to be loaded", the LSTM would learn to focus on the image region corresponding to the "airplane". It considers the sequence generated so far and focuses on the part of the image that needs to be described next, which is exactly what the attention mechanism does. Specifically, we use soft Attention, where the pixel weights sum to 1. For feature maps with a spatial resolution of S, the attention weight α at timestep δ satisfies the following equation:

$$\sum_{s=1}^{S} \alpha_{s,\delta} = 1 (1)$$

At each generation step, the feature maps and the previous hidden state are used to compute the attention weights for each pixel in the Attention network. The previously generated word and the attention-weighted feature maps are fed into the LSTM to generate the next word. Specifically, the output h of the LSTM at the current timestep is used as the semantic feature for word prediction. The prediction score for each word in the dictionary is calculated using the following equation:

$$z = w_d \cdot (h \otimes r) + b_d (2)$$

where $r \sim Bernoulli(p)$ is the dropout mask with $p = 0.5$, and $\otimes$ represents the Hadamard product of matrices. $w_d \in \mathbb{R}^{1\times D}$ denotes the weight, and b_d is a scalar bias. Finally, we obtain the prediction scores $z = [z_1, z_2, \cdots, z_D]$, representing the scores for the D words in the dictionary.

3.4. Activity prediction

Attention plays a crucial role in the HVS and has been successfully applied to DNNs [41, 42]. In Section 3.3 Image captioning, attention is used to assist the LSTM in determining which specific feature maps should be focused on for word prediction. Here, we employ channel attention to prioritize important feature maps. Given the global information $Fm = AvgPool(\boldsymbol{FM}) \in \mathbb{R}^{1\times C}$ of all channels, it is further formulated as follows:

$$ca = \sigma(\boldsymbol{W}_{ca} \cdot Fm + b_{ca}) (3)$$

where $\boldsymbol{W}_{ca} \in \mathbb{R}^{C\times C}$ and $b_{ca} \in \mathbb{R}^{1\times C}$ are the learned weight matrix and bias term, respectively. σ denotes the sigmoid function. Finally, the channel attention $ca \in \mathbb{R}^{1\times C}$ is applied to the original feature map, $ca \otimes \boldsymbol{FM}$, where $\otimes$ denotes the Hadamard product.

Furthermore, the population activity of a single voxel encodes features within limited and contiguous regions of the visual field [43, 44]. For the flexible RF, each value is a randomly initialized, learnable independent parameter. As the RFs of the high-level visual cortices expand, the positional sensitivity of voxels to visual stimuli gradually decreases, with increasing focus on global features and semantic information. It may not learn a meaningful spatial representation through the traditional RF model. In biological visual systems, RFs not only receive visual information, but are also modulated by attention, emphasizing visual areas relevant to the current cognitive task. To address this, we have innovatively designed an RF module constrained by visual attention, termed "Atten-RF":

$$rf'_v = \alpha' \otimes rf_v (4)$$

where rf_v is the flexible RF of voxel v, $\alpha' \in \mathbb{R}^S$ represents the mean of the attention α across all timesteps, and $\otimes$ denotes the Hadamard product. Attention is used as the weight of the RF, allowing it to both encode the original spatial representation and capture the visual semantic information relevant to the ICT. The feature maps are mapped to low-dimensional feature representations through the Atten-RF as follows:

$$f_v = g_{out}\left(g_{in}(\boldsymbol{FM}) \otimes rf'_v\right) (5)$$

Additionally, following previous research [11], a fully differentiable nonlinearity is applied before and after spatial pooling:

$$g_{in}(\cdot) = g_{out}(\cdot) = tanh(\cdot)\, log\left(1 + |\cdot|\right)(6)$$

In terms of mapping from feature space to voxel activity space, nonlinear transformation is more appropriate than linear transformation. However, when only a small amount of high-dimensional fMRI data is available for training, there is a risk of overfitting [45]. Previous research has demonstrated the excellent performance of linear models in neural encoding [46, 47], where the activity pattern of the v-th voxel, denoted as $\bar{y}_v$, is predicted through a linear combination of joint features f_v:

$$\hat{y}_v = w_v \cdot f_v + b_v(7)$$

where $w_v \in \mathbb{R}^{1\times K}$ denotes the weight, and b_v is a scalar bias specific to the voxel.

3.5. Multi-domain joint optimization

To achieve accurate encoding of voxel activities and ensure rapid convergence of the model, we designed a multi-domain joint loss to train our model, as illustrated in Figure 2.

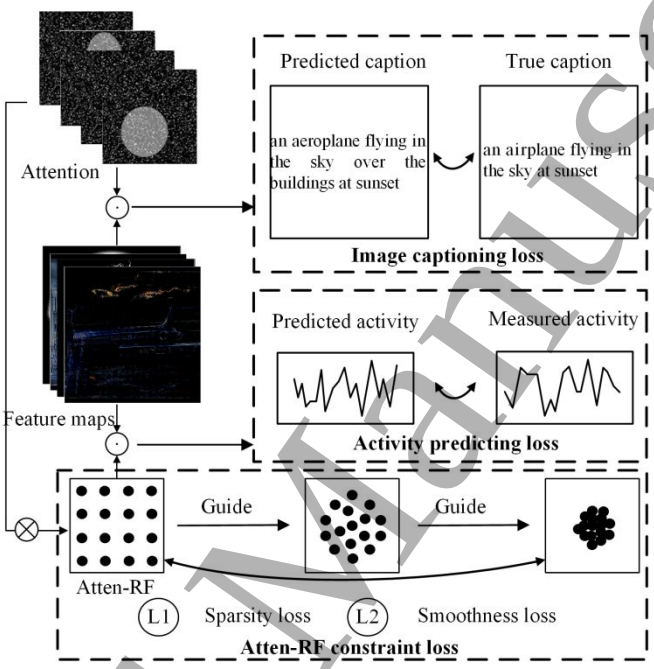

Figure 2. The schematic diagram of the multi-domain loss, which comprises the image captioning loss, activity predicting loss, and RF constraint loss (indicated by the dashed rectangle).

Image captioning loss. The prediction score for each word in each sample is denoted as $z = [z_1, z_2, \cdots, z_D]$, which is converted into the probability distribution $\hat{t}_d$ as follows:

$$\hat{t}_d = \frac{\exp(z_d)}{\sum_{i=1}^{D} \exp(z_i)}(8)$$

The predicted probability corresponding to the true label t_d is selected, and the negative log-likelihood loss is computed as follows:

$$L_{IC} = -\log\left(\hat{t}_d, t_d\right)(9)$$

Finally, we average the losses across the batch of samples to obtain the image captioning loss $\mathcal{L}_{\mathrm{IC}}$.

Activity predicting loss. The traditional voxel-wise encoding approach involves training a separate model for each voxel, leading to low encoding efficiency, particularly when the number of voxels ranges from 10^3 to 10^5. The mean squared error (MSE) loss between the predicted and true voxel activities in a specific ROI, which is the most commonly used loss function for encoding model training, is defined as follows:

$$L_{AP} = \frac{1}{V}\sum_{v=1}^{V}(y_v - \hat{y}_v)^2(10)$$

Finally, the activity prediction loss $\mathcal{L}_{\mathrm{AP}}$ is obtained by averaging across the batch of samples.

Atten-RF constraint loss. The general RF rf is weighted by attention to produce the Atten-RF rf'. L1 regularization is applied to enforce sparsity in the Atten-RF, while L2 regularization on the Laplacian operator ensures smoothness, enhancing the local structure and interpretability of the Atten-RF. The loss is defined as follows:

$$L_{RF} = \frac{1}{S}\sum_{s=1}^{S} \|rf'_s\|_1 + \frac{1}{S}\sum_{s=1}^{S} \|laplacian_{rf'_s}\|_2 (11)$$

where S is the spatial resolution of the Atten-RF. The Laplacian operator, applied to the Atten-RF, is computed via a convolution operation as follows:

$$laplacian_{rf'_s} = rf'_s * \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} (12)$$

Finally, the Atten-RF constraint loss $\boldsymbol{\mathcal{L}}_{RF}$ is obtained by averaging across all voxels.

The multi-domain joint loss integrates $\boldsymbol{\mathcal{L}}_{IC}$, $\boldsymbol{\mathcal{L}}_{AP}$, and $\boldsymbol{\mathcal{L}}_{RF}$:

$$\boldsymbol{\mathfrak{L}} = \boldsymbol{\mathcal{L}}_{IC} + \boldsymbol{\mathcal{L}}_{AP} + \boldsymbol{\mathcal{L}}_{RF} (13)$$

During training, the image captioning task and voxel encoding process are jointly optimized to learned shared visual representations and patterns. This collaborative learning approach helps prevent overfitting in single task/process and promotes the development of more generalized and robust representations. The complete pseudocode for training the proposed model is presented in Algorithm 1.

Algorithm 1: Pseudocode for training the parameters of our model

Input: training dataset $\{\boldsymbol{X},\boldsymbol{T},\boldsymbol{Y}\}$; feature map extractor $\boldsymbol{\phi}$; image caption generator $\boldsymbol{\psi}$; activity prediction generator $\boldsymbol{\varphi}$; tunable parameters $\theta_{\boldsymbol{\phi}}$, $\theta_{\boldsymbol{\psi}}$, $\theta_{\boldsymbol{\varphi}}$; optimizer $Adam$; epochs; number of batches n_B.

Randomly initialize model parameters $\theta_{\boldsymbol{\phi}}$, $\theta_{\boldsymbol{\psi}}$, $\theta_{\boldsymbol{\varphi}}$.

For e in epochs:

  // Training

  Randomly shuffle the dataset $\{\boldsymbol{X},\boldsymbol{T},\boldsymbol{Y}\}$

  For $i = 1,\cdots,n_B$:

    Take a batch set $\{\boldsymbol{X}_i,\boldsymbol{T}_i,\boldsymbol{Y}_i\}$

    Feature map extraction: $\boldsymbol{FM}_i \leftarrow \boldsymbol{\phi}(\boldsymbol{X}_i)$

    Image captioning: $\hat{\boldsymbol{T}}_i \leftarrow \boldsymbol{\psi}(\boldsymbol{FM}_i)$

    Image captioning loss: $\boldsymbol{\mathcal{L}}_{\mathrm{IC}} \leftarrow loss(\hat{\boldsymbol{T}}_i,\boldsymbol{T}_i)$

    Activity predicting: $\hat{\boldsymbol{Y}}_i \leftarrow \boldsymbol{\varphi}(\boldsymbol{FM}_i)$

    Activity predicting loss: $\boldsymbol{\mathcal{L}}_{AP} \leftarrow loss(\hat{\boldsymbol{Y}}_i,\boldsymbol{Y}_i)$

    Joint loss: $\boldsymbol{\mathfrak{L}} \leftarrow \boldsymbol{\mathcal{L}}_{IC} + \boldsymbol{\mathcal{L}}_{AP} + \boldsymbol{\mathcal{L}}_{RF}$

    Update $\theta_{\boldsymbol{\phi}}, \theta_{\boldsymbol{\psi}}, \theta_{\boldsymbol{\varphi}} \leftarrow Adam(\boldsymbol{\mathfrak{L}};\theta_{\boldsymbol{\phi}}, \theta_{\boldsymbol{\psi}}, \theta_{\boldsymbol{\varphi}})$

  // Validation

  Repeat the above steps using the validation set

  Multi-domain joint loss: $\boldsymbol{\mathfrak{L}}_{val}$

  While $\boldsymbol{\mathfrak{L}}_{val}$ increases for 3 consecutive times:

    Break

Output: Optimal network parameters $\theta^*_{\boldsymbol{\phi}}$, $\theta^*_{\boldsymbol{\psi}}$, $\theta^*_{\boldsymbol{\varphi}}$.

4. Experiments

4.1. Dataset and preprocessing

Natural Scenes Dataset (NSD). It collected fMRI data from eight subjects as they viewed images captured from natural scenes [48]. In our study, we utilized data from 4 subjects (sub1, sub2, sub5, sub7) who completed all trials [49]. During the training phase, 8,859 unique images were presented to each subject, with each image displayed for 3 seconds, resulting in 24,980 fMRI trials (with up to 3 repetitions per image). During the testing phase, 982 images were presented, yielding 2,770 fMRI trials. These 982 testing images were shared across subjects, ensuring consistency in cross-subject comparisons. All fMRI data underwent two preprocessing steps: temporal and spatial interpolation to correct slice timing differences and head motion, and generation of Z-score normalized single-trial beta estimates using the GLMSingle method, as detailed in [48]. To enhance the signal-to-noise ratio, multiple fMRI trails of repeated images were averaged. The NSD dataset delineates multiple ROIs. For this experiment, we selected 7 high-level visual ROIs: Occipital Face Area (OFA), Fusiform Face Area (FFA), Occipital Word Form Area (OWFA), Visual Word Form Area (VWFA), Occipital Place Area (OPA), Extrastriate Body Area (EBA), and Fusiform Body Area (FBA). The number of voxels for each selected ROI is presented in Table 1 (The schematic diagram is shown in Figure 3A). For the visual stimuli, we downsampled the natural images from a size of 425×425 to 256×256 to reduce computational complexity. Since the stimulus images in the NSD dataset are derived from the COCO dataset, which includes descriptive text for each image provided by 5 annotators, we randomly select one text description for

each image as its corresponding label. Additionally, we used 10% of the training samples as validation data to monitor the training progress of our model.

Stability selection of voxels. Due to individual differences in brain representation, brain activity patterns can vary from trial to trial, even for the same visual stimulus. To enhance the stability of neural encoding, we used a stability selection strategy for the NSD dataset [50]. We selected voxels with high consistency in activation patterns across trials of the same visual stimulus. Specifically, the Pearson correlation coefficient (PCC) was used to assess the consistency of activation patterns across trials:

$$r = \frac{\sum(y - \bar{y})\,(y' - \bar{y}')}{\sqrt{\sum (y - \bar{y})^2} \cdot \sqrt{\sum (y' - \bar{y}')^2}} (14)$$

where y and y' represent the fMRI signals from two different trials under the same visual stimulus. The diagram illustrating the stability selection process is shown in Figure 3B. Since each subject had 1-3 trials for each stimulus image, we uniformly selected images with 3 trials for statistical analysis. Instead of applying a fixed PCC threshold, voxels were ranked by PCC values, and the top N voxels were selected per ROI, with N determined by the ROI with the fewest voxels in each subject. This ranking-based selection achieves a similar effect to applying a PCC threshold, but avoids inconsistencies in voxel counts caused by differences in brain regions. Specifically, the number of voxels selected for each ROI of subjects 1, 2, 5, and 7 was 355, 441, 438, and 316, respectively.

Construction of dictionary. Captions serve as both the target and the input for the LSTM, with each generated word used to predict the next one. To generate the first word, we introduce a special token <start>, as the zeroth word. The last word is predicted as <end>, indicating to the LSTM when to terminate the ICT. Since captions vary in length, we pad them with <pad> tokens to ensure they are all the same fixed size when passed through the model. In summary, we constructed a dictionary that maps each word in the corpus to an index, including the special tokens <start>, <end>, and <pad>.

Table 1. For each subject, voxel counts were obtained across seven high-level visual regions (OFA, FFA, OWFA, VWFA, OPA, EBA, and FBA). The region with the fewest voxels is highlighted in bold and used as the subject-specific threshold for voxel stability selection.

Subs	OFA	FFA	OWFA	VWFA	OPA	EBA	FBA
Sub1	**355**	794	464	1083	1611	2971	826
Sub2	**441**	869	519	821	1381	3439	1217
Sub5	782	907	**438**	941	1332	4587	968
Sub7	**316**	484	628	465	1083	3062	552

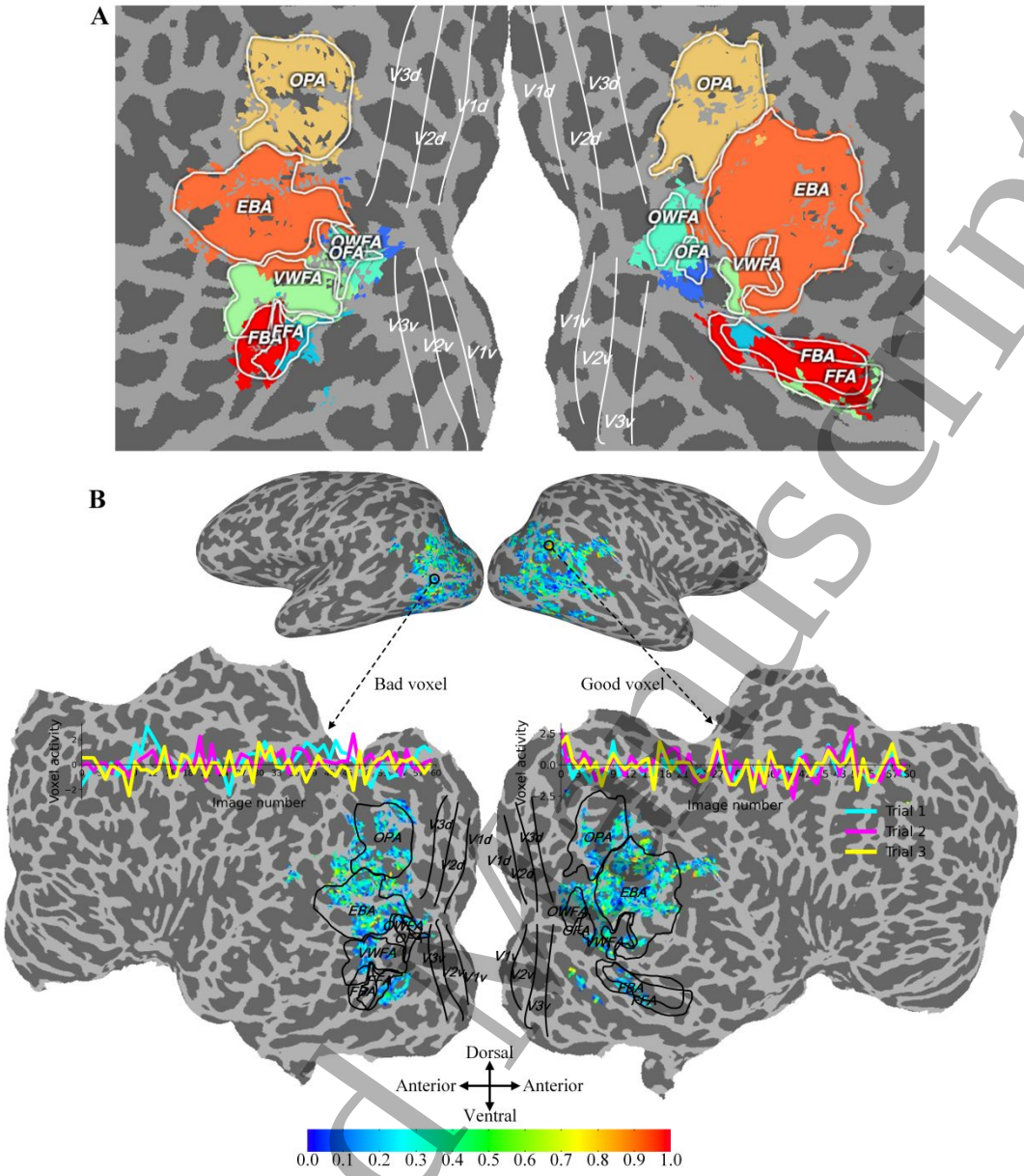

Figure 3. Data and preprocessing diagram. (A) The schematic diagram of the selected ROIs in the NSD dataset. (B) The voxel stability map on the flat surface of the brain for subject 1. The stability score of each voxel is quantified as the average PCC across all pairwise combinations of the three trials.

4.2. Performance evaluation metrics

To quantitatively evaluate encoding performance based on different properties, we used three standard similarity metrics: MSE, PCC, and coefficient of determination R^2. While MSE focuses on point-to-point prediction accuracy, PCC and R^2 capture variations in texture and the overall goodness of fit, which are particularly significant in neuroscience [25]. The MSE_v, $\mathrm{PCC_v}$ and R_v^2 for the v-th voxel are calculated as follows:

$$MSE_v = \frac{1}{N}\sum_{n=1}^{N}\left(y_v^n - \hat{y}_v^n\right)^2 \quad (15)$$

$$PCC_v = \frac{\sum_{n=1}^{N}(y_v^n - \bar{y}_v)(\hat{y}_v^n - \bar{\hat{y}}_v)}{\sqrt{\sum_{n=1}^{N}(y_v^n - \bar{y}_v)^2}\cdot\sqrt{\sum_{n=1}^{N}(\hat{y}_v^n - \bar{\hat{y}}_v)^2}} \quad (16)$$

$$R_{\hat{v}}^2 = 1 - \frac{\sum_{n=1}^{N} \left(y_v^n - \hat{y}_v^n\right)^2}{\sum_{n=1}^{N} \left(y_v^n - \bar{y}_v\right)^2} (17)$$

where $\bar{y}_v$ and $\bar{\hat{y}}_v$ are the mean measured and predicted values of the entire test set for the *v*-th voxel.

In addition, we also use the top-5 classification accuracy (Top5-ACC) metric to further evaluate the performance of our model in the ICT.

4.3. Performance comparison

We propose an end-to-end neural encoding model based on the ICT. The following five advanced methods are used for comparison:

1) Gabor-fwRF (2018). It first extracts Gabor wavelet feature maps and calculates the weighted sum within the spatial extent of a 2-*D* Gaussian RF (for details, refer to [20]). Finally, it regresses all features simultaneously to predict the brain response with high accuracy.

2) AlexNet-gpf (2022). It first extracts feature maps from a pre-trained AlexNet, and then a 2-D Gaussian RF is applied to further extract features. A voxel-wise regression model is subsequently constructed to predict brain activity [48].

3) GNet-fpf (2023). It extracts feature maps from GNet, which are then passed through a flexible pooling field for further feature extraction. Linear layers are used for predicting voxel activity [11].

4) AlexNet-fpf (2023). It initially extracts feature maps from pre-trained AlexNet, followed using a flexible pooling field for further feature extraction. A voxel-wise regression model is then built to predict brain activity [11].

5) ResNet101-fpf. Unlike AlexNet-fpf, this method extracts feature maps from a pre-trained ResNet101 and fine-tunes its final residual block during end-to-end training. In contrast to our approach, it does not incorporate the ICT component.

To ensure fair comparison, all competing methods followed the hyperparameter settings recommended in their original papers. Additionally, key hyperparameters such as learning rate, batch size, and regularization coefficients were further tuned on the validation set using a grid search strategy. All models were trained and evaluated under the same training, validation, and test splits, where the validation set was solely used for hyperparameter selection and the test set was strictly reserved for performance reporting. This procedure guarantees that each competing model operates under its optimal and unbiased performance.

4.4. Implementation details

During the training phase, the parameters of the pre-trained ResNet101 are either kept frozen or selectively fine-tuned. In particular, we fine-tune only the final residual block due to its closer relevance to semantic encoding, while the rest of the model's parameters are learned during training. The dimensions of the word embeddings, attention linear layers, and LSTM are set to 512, with the maximum caption length for image captioning set to 40. The number of epochs and batch size were set to 50 and 128, respectively. The Adam optimizer with an initial learning rate of 0.001 was used. When the loss on the validation set increases, the learning rate is decayed by a factor of 0.8. If the loss increases for 3 consecutive epochs, early stopping is applied. Gradient clipping is performed during backpropagation to prevent gradient explosion. To minimize the effect of randomness, we repeated the experiment 5 times and calculated the average for each subject. All experiments were conducted on a workstation equipped with a 12th Gen Intel (R) Core (TM) i7-12700K CPU and an NVIDIA GeForce RTX 3090 GPU. The neural network models were implemented using the publicly available Pytorch framework.

4.5. Experimental results

1) Quantitative analysis of encoding performance. Table 2 presents the quantitative comparison between the proposed model and the advanced methods discussed. The results are reported as the $mean \pm std$ deviation across five random seeds and four subjects, with the best performance in each ROI highlighted. Overall, except for slightly inferior MSE and PCC values in the OWFA region compared to ResNet101-fpf, our model consistently outperforms existing advanced methods in neural encoding performance. Specifically, in terms of the average MSE (lower is better) across seven high-level visual ROIs, our model surpasses all other methods: 0.122 lower than Gabor-fwRF, 0.080 lower than AlexNet-gpf, 0.047 lower than GNet-fpf, 0.031 lower than AlexNet-fpf, and 0.025 lower than ResNet101-fpf. Regarding the average PCC metric (higher is better) over the same ROIs, our model also outperforms the others: 0.178 higher than Gabor-fwRF, 0.141 higher than AlexNet-gpf, 0.052 higher than GNet-fpf, 0.032 higher than AlexNet-fpf, and 0.024 higher than ResNet101-fpf. As for the average R^2 metric (higher is better), our model again outperforms the others: 0.155 higher than Gabor-fwRF, 0.148 higher than AlexNet-gpf, 0.091 higher than GNet-fpf, 0.065 higher than AlexNet-fpf, and 0.055

higher than ResNet101-fpf. For statistical analysis, we employed paired two-tailed t-tests to compare the performance of our model with each advanced method under the same subject and ROI conditions. The significance threshold was set to $p < 0.05$, where * indicates significant improvement ($p < 0.05$), ** indicates highly significant improvement ($p < 0.005$), and – indicates no significant difference ($p \geq 0.05$). All tests were conducted across five random seeds and four subjects to ensure robustness of the statistical conclusions. The statistical analysis results of our method and the comparative methods are also shown in Table 2.

The methods that rely on the strong assumption of Gaussian RF (-fwRF, -gpf) are less effective at fitting the encoding range and size of voxels in the high-level visual cortices. In contrast, the improved methods based on weaker assumptions and learnable flexible RF (-fpf) significantly improve the encoding performance of the voxels. Furthermore, the brain-optimized network [11] (GNet-fpf) exhibits inferior performance compared to the task-optimized network [11] (AlexNet-fpf). We attribute this to the scarcity of {fMRI, image} sample pairs, which poses a major challenge for training models from scratch to directly predict voxel responses, particularly in high-level visual areas responsible for encoding complex and abstract visual features. Compared to a standalone neural encoding process, our model significantly improves encoding performance across high-level visual brain regions by integrating the ICT.

2) Comparison of encoding performance PCC. For a single voxel, a PCC value greater than 0.27 between the predicted and measured response is statistically significant compared to its null hypothesis distribution ($p < 0.001$) [39, 51]. Taking subject 1 as an example, we conducted a statistical analysis of the PCC values across all voxels within each ROI and compared the results with those obtained using the ResNet101-fpf method. As shown in Figure 4A, our model outperforms ResNet101-fpf across all seven ROIs, with comparable performance observed only in the OFA and OWFA. On the other hand, the PCC values of all voxels exhibit a clear linear relationship between the two methods, indicating that the integration of ICT and joint training consistently improves prediction performance of voxel-wise responses. Furthermore, the results in Figure 4B demonstrate that for ROIs with strong encoding capabilities, which contain fewer voxels with PCC < 0.27, such as the FFA, VWFA, OPA, EBA, and FBA, our model achieves significant performance gains. In contrast, even for ROIs with weaker encoding capabilities, which are those containing more voxels with PCC < 0.27, such as the OFA and OWFA, our model still achieves noticeable improvements in encoding performance. These findings clearly demonstrate the strong generalization ability and robustness of our model in predicting voxel responses across ROIs of different encoding abilities.

To assess the consistency of encodability across subjects, we projected the PCC of all voxels from the four subjects onto the flat surface space for visualization and calculated the average PCC for each region, as shown in Figure 5. The distribution of encoding performance across each ROI is relatively consistent among subjects, highlighting the generalization and robustness of our model. On the other hand, neuroscience research indicates that the functions and information encoding mechanisms of specific ROIs exhibit high consistency across subjects. For instance, the EBA is sensitive to body shape information in images, while the OPA excels at encoding global scene and spatial relationships. This demonstrates that our model effectively captures the functional patterns of each ROI.

3) Comparison of explainable unique variances. To compare the independent contributions of visual features extracted by our proposed model (A) and ResNet101-fpf model (B) to voxel activity prediction, we first trained voxel-wise ridge regression mappings from the visual features of models A and B to voxel responses, yielding explained variances R_A^2 and R_B^2, respectively. Subsequently, a ridge regression mapping was trained using the concatenated visual features from both models, resulting in a combined explained variance R_{AB}^2. Based on these values, the unique variance contributions of models A and B can be expressed as:

$$Unique(A) = R_{AB}^2 - R_B^2$$
$$Unique(B) = R_{AB}^2 - R_A^2 \quad (18)$$

This metric quantifies the unique contribution of each model to the explained variance [29], isolating the proportion of variance accounted for by one model after removing the influence of the other. By computing these unique variance values, we aim to assess the relative effectiveness of our model and ResNet101-fpf in modeling the neural encoding process, and indirectly evaluate the comprehensiveness and accuracy of the visual information extracted by each model. As shown in Figure 6A and Figure 6B, the comparison results reveal a clear difference in the unique contribution of the two models. To provide a more intuitive visualization, we computed the difference $\Delta U = Unique(A) - Unique(B)$, and projected the results onto the corresponding 3D and flattened cortical coordinates, as illustrated in Figure 6C. These findings consistently indicate that the visual features learned by our model contribute more significantly to voxel activity prediction compared to those extracted by ResNet101-fpf. We attribute this improvement to the guidance and alignment provided by the ICT, which not only enhances the

model's ability to learn more comprehensive semantic representations but also improves the alignment between these representations and brain activity in high-level visual cortex.

Table 2. Comparison of the encoding performance between our model and advanced methods. The results are shown in the form of the $mean \pm std$ of five random seeds and four subjects, with the best performance in each ROI highlighted in bold. A paired two-tailed t-test was employed to compare the performance of our model with each advanced method under the same subject and ROI conditions. The significance threshold was set to $p < 0.05$, where * indicates significant improvement ($p < 0.05$), ** indicates highly significant improvement ($p < 0.005$), and – indicates no significant difference ($p \geq 0.05$).

Metrics	Algorithms	OFA	FFA	OWFA	VWFA	OPA	EBA	FBA
MSE↓	Gabor-fwRF	.962±.003**	.897±.002**	.966±.003**	.902±.003**	.895±.002**	.711±.003**	.879±.003**
	AlexNet-gpf	.931±.005**	.859±.004**	.934±.006**	.868±.003**	.853±.002**	.631±.004**	.836±.002**
	GNet-fpf	.887±.004**	.807±.004**	.919±.005–	.814±.003**	.790±.002**	.664±.005**	.805±.002**
	AlexNet-fpf	.894±.004**	.798±.003**	.923±.003**	.803±.001**	.771±.003**	.591±.004**	.794±.003**
	ResNet101-fpf	.889±.006**	.780±.003**	**.918±.003**–	.791±.003**	.762±.005**	.607±.002**	.782±.001**
	Ours	**.875±.003**	**.764±.001**	.920±.006	**.777±.002**	**.749±.002**	**.501±.003**	**.769±.003**
PCC↑	Gabor-fwRF	.186±.003**	.234±.002**	.147±.001**	.262±.003**	.339±.001**	.423±.005**	.268±.005**
	AlexNet-gpf	.203±.003**	.266±.003**	.185±.002**	.299±.001**	.379±.003**	.461±.004**	.324±.001**
	GNet-fpf	.284±.005**	.389±.001**	.249±.004**	.385±.002**	.464±.001**	.577±.007**	.392±.004**
	AlexNet-fpf	.286±.002**	.405±.004**	.254±.003**	.405±.002**	.487±.002**	.631±.003**	.410±.002**
	ResNet101-fpf	.292±.004**	.405±.003**	**.271±.002**–	.411±.003**	.488±.005**	.645±.003**	.426±.003**
	Ours	**.308±.001**	**.439±.002**	.270±.004	**.437±.002**	**.513±.002**	**.696±.002**	**.441±.002**
R^2↑	Gabor-fwRF	.053±.004**	.061±.004**	.044±.004**	.077±.001**	.108±.003**	.211±.005**	.076±.003**
	AlexNet-gpf	.055±.001**	.075±.001**	.052±.003**	.103±.003**	.122±.005**	.251±.001**	.023±.005**
	GNet-fpf	.096±.003**	.112±.002**	.065±.005**	.124±.005**	.251±.003**	.302±.003**	.132±.004**
	AlexNet-fpf	.096±.004**	.147±.004**	.069±.001**	.148±.004**	.270±.001**	.376±.002**	.154±.001**
	ResNet101-fpf	.101±.005**	.147±.005**	**.083±**.002–	.156±.004**	.269±.002**	.399±.004**	.175±.003**
	Ours	**.114±**.001	**.219±**.002	**.083±**.003	**.218±**.005	**.387±**.002	**.474±**.005	**.221±**.002

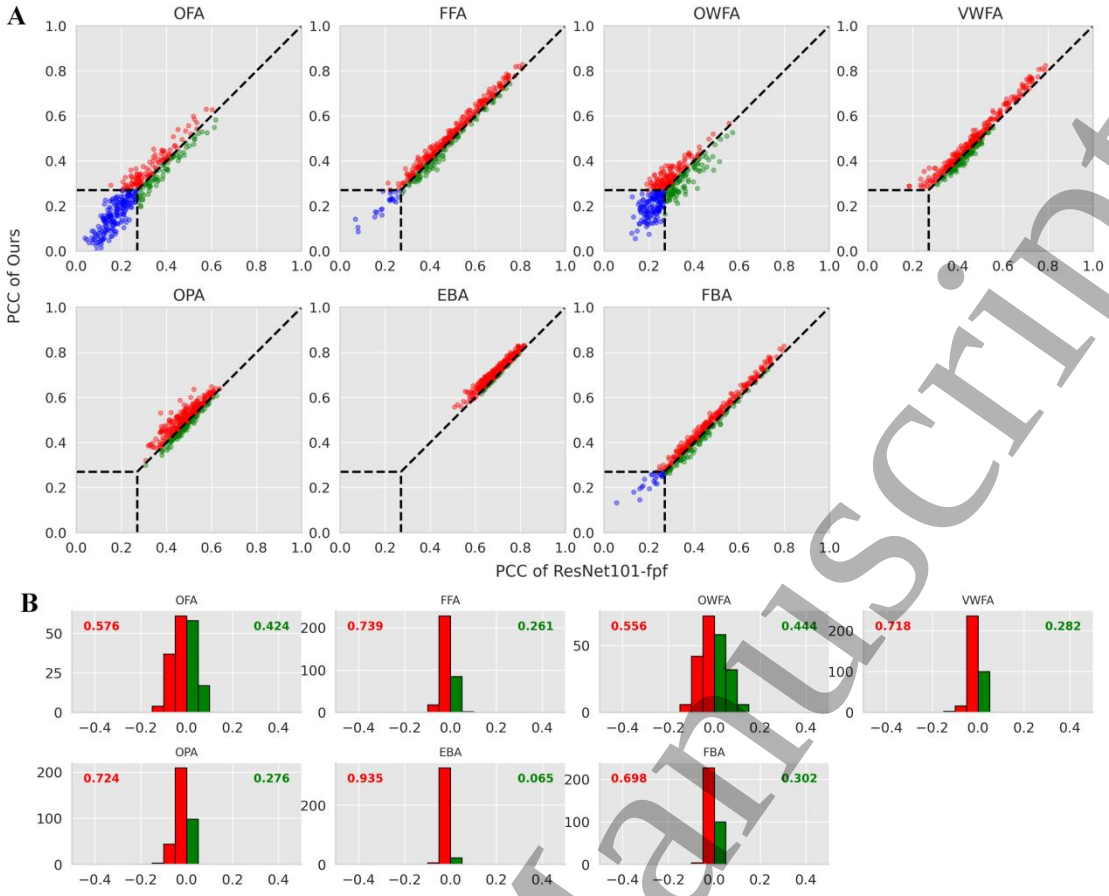

Figure 4. Comparison of encoding performance between our method and ResNet101-fpf on Sub 1. (A) Comparison of PCC, where red voxels represent better performance by our model, green voxels indicate better performance by ResNet101-fpf, and blue voxels represent PCC<0.27. (B) Comparison of the proportion of superior voxels, where red bars indicate our model and green bar present ResNet101-fpf.

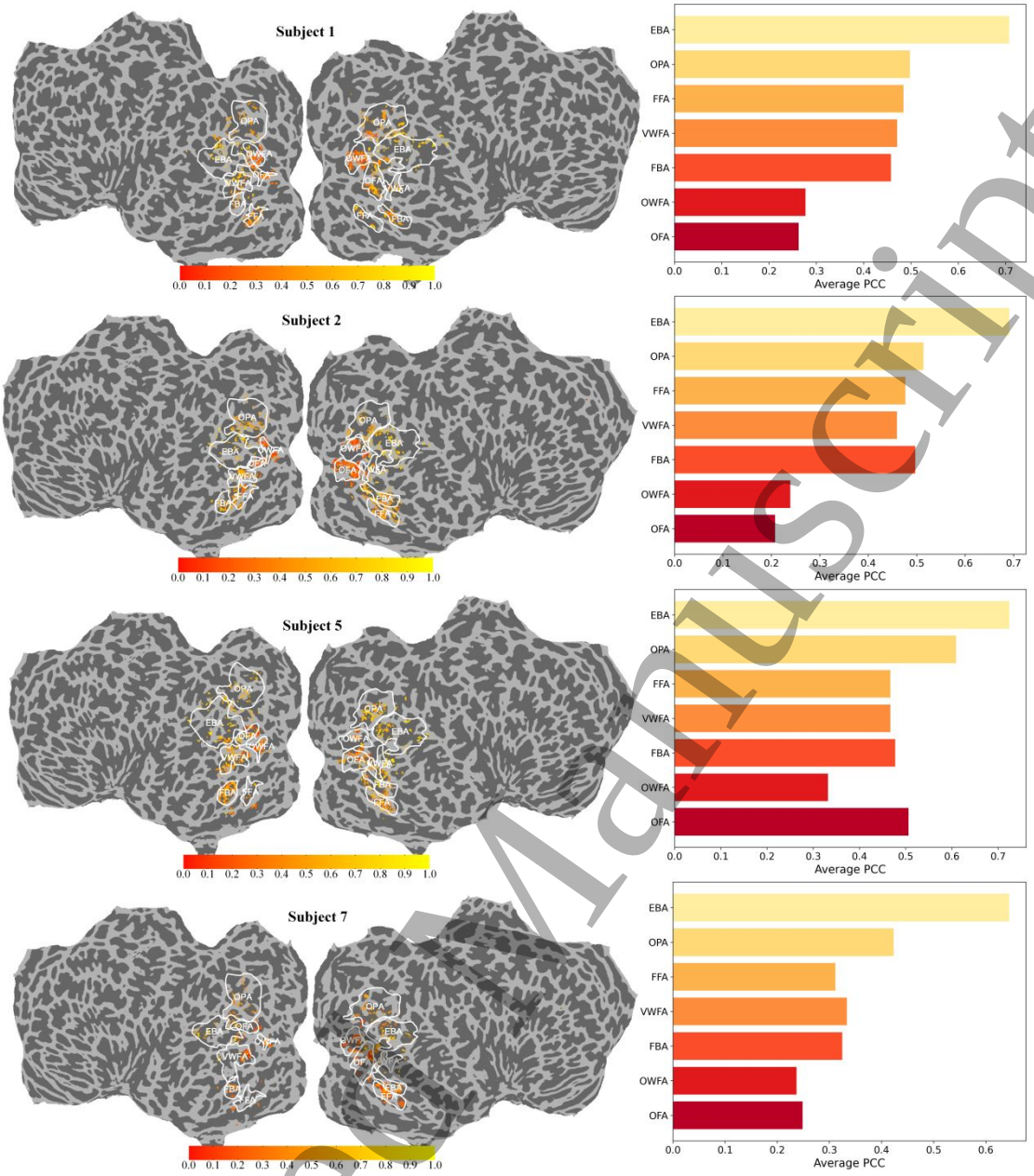

Figure 5. The left panel shows the projection of voxel-wise PCC values onto the cortical flat map of each subject constructed using Pycortex [52], while the right panel presents the mean PCC values across all voxels within each ROI. All results are normalized to the range of 0–1 and displayed in the same order.

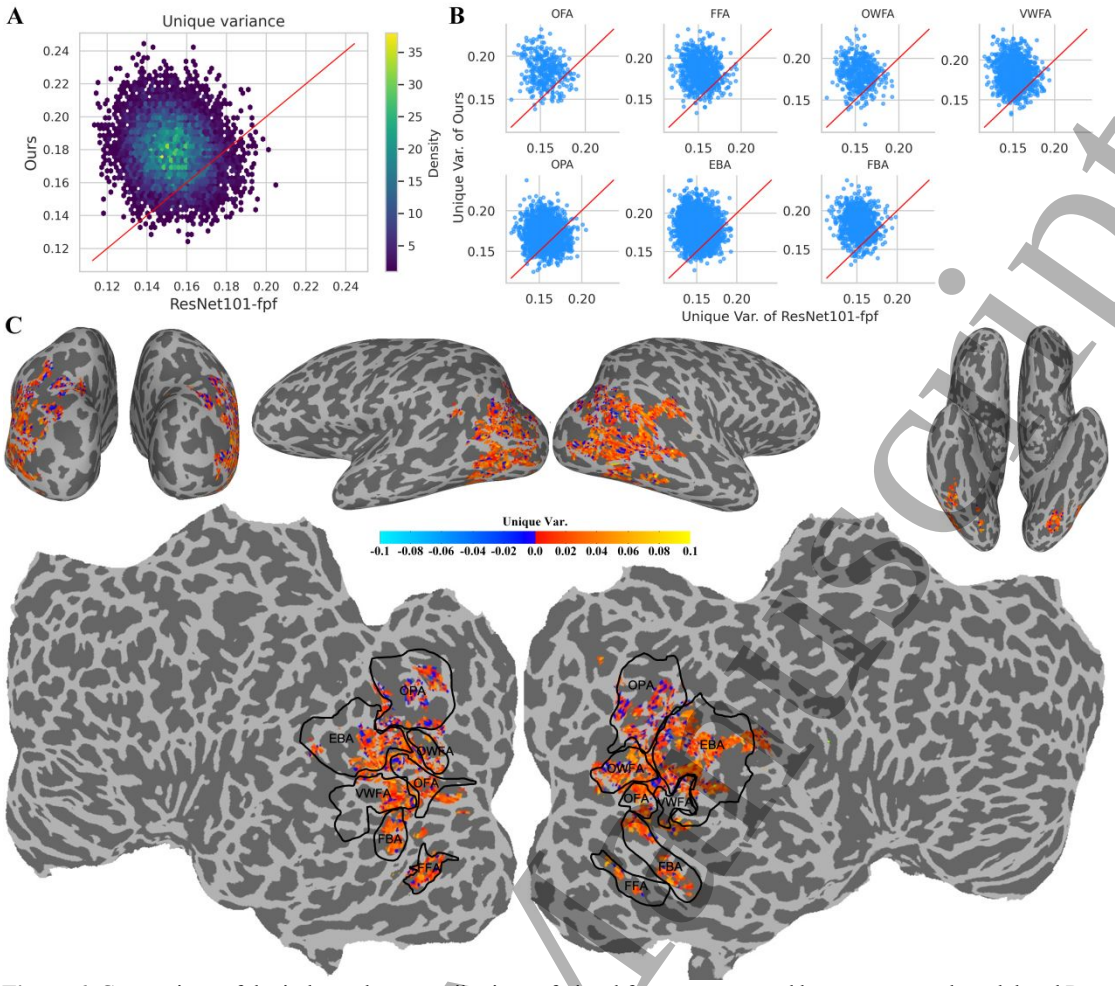

Figure 6. Comparison of the independent contributions of visual features extracted by our proposed model and ResNet101-fpf to voxel activity prediction. (A) Scatter density plot of the unique variance for the two neural encoding models: the x-axis represents the unique variance of ResNet101-fpf, and the y-axis represents that of ours. (B) Scatter plot of the unique variance distribution across different brain regions. (C) Projection of ΔU onto the visual cortex, visualized on flattened, lateral, posterior and bottom views.

4) Visualization of attention and RF. To verify our model's ability to learn attention, we intuitively visualized the attention changes in the EBA of subject 1 in response to the image description. We resized the 14×14 sparse attention map to 256×256 and overlaid it on the original stimulus image, as shown in Figure 7A. We observed that the area of attention shifted from the "table" to the "chair", and then to the "portrait" and "fireplace" as the image description progressed. This clearly demonstrates that the attention mechanism dynamically captures the correspondence between visual information and its linguistic description. At the same time, this shift in attention highlights the model's ability to focus on key areas relevant to the current semantics, further validating its interpretability in semantic decoding tasks. This also supports the enhancement of neural encoding performance by integrating the ICT, providing a solid foundation for its efficacy.

It is an interesting open question whether the features learned by task-optimized networks like AlexNet are similar to or diverge from those learned by brain-optimized networks like GNet [48]. To bridge the domain gap between the ICT (task-optimized network) and the visual neural encoding process (brain-optimized network), we incorporated an Atten-RF module to enhance the effectiveness of multi-domain joint training. Specifically, the ICT utilizes the attention mechanism to focus on key areas of the input image, capturing representations related to the semantic description. Meanwhile, the visual neural encoding process models the spatial representation of the visual cortex using the RF module, revealing how the brain processes and encodes visual information. The Atten-RF we designed not only offers a novel perspective on understanding visual encoding but also contributes to advancing the development of more biologically inspired models for visual information processing. We compared the roles and performances of the "Attention" module and the "RF" module at different voxels from distinct brain regions, as shown in Figure 7B. Attention reflects the salient areas that the model focuses on in the ICT, typically corresponding to the key objects or semantics in the image. The highlighted areas of the average

attention (second column) are concentrated on key objects or regions with high semantic information in the image, such as people, furniture, streetlights, etc. This demonstrates that the attention mechanism effectively captures the salient features relevant to the task. Subsequently, for each brain region, we selected the voxel with the highest PCC, extracted its 14×14 RF weight map, interpolated it to 256×256, and overlaid it onto the original image to visualize the spatial distribution of RFs across regions (third to ninth columns). We found that the RFs of the most predictive voxels across all regions are spatially close and highly consistent with the hotspot areas in the averaged attention maps. This observation suggests that high-level visual areas may share partially similar spatial attention patterns when processing semantic information, which is consistent with prior cognitive neuroscience studies proposing common feature extraction mechanisms in higher visual cortex [53]. At the same time, the similarity observed across ROIs could also reflect the convergence of distributed global feature representations that yield the highest predictive performance. Such convergence has also been reported in deep visual networks, where later-layer features often become highly similar and encode global patterns rather than explicit local selectivity, as noted in DINO v2 [54]. We therefore consider that RF visualizations may not solely indicate shared attention mechanisms but could also reflect the dominance of global representations in high-level areas. For example, in the OFA region, for the third-row human stimuli, the RF hotspot nearly coincides with the high-response area of the averaged attention map, supporting the central role of OFA in face and body feature processing and suggesting its potential coupling with global attention mechanisms. These interpretations are not mutually exclusive and together point to the complexity of semantic processing in high-level visual cortex. While convergence toward global feature representations may explain the overall similarity observed across ROIs, it does not preclude the existence of category-specific preferences in certain regions. Indeed, some brain regions exhibit more focused RF distributions for specific categories of stimuli, such as the FFA for person-related stimuli and the EBA for furniture-related stimuli, which may reflect category-specific preferences in semantic perception processing. This finding also indicates that the spatial attention patterns learned by ICT partially guide the RF-based neural encoding process.

5) Visualization of semantic encoding in different ROIs. Through the ICT, we analyzed the differences in how various high-level visual cortices represent visual stimuli. We used semantic vocabulary to describe and explain the functional characteristics of these cortices. We counted the word frequencies of the predicted texts for all test images and normalized them relative to the word frequencies of the real texts (Figure 8(a)). Using a word visualization tool (Wordcloud), we then illustrated the semantic word distribution of specific ROIs. The more frequently a semantic word appears, the larger the area it occupies in the image space. Different words are represented in different colors. We show the semantic word distribution for three representative ROIs: EBA, OPA, and OWFA, which are shown in Figure 8(b), (c), and (d), respectively. Cognitive neuroscience research indicates that the EBA is a specific ROI associated with the visual perception of body shapes and body parts. As shown in Figure 8(b), it encodes semantics related to actions such as 'walking', 'sitting', and others. The OPA, a ROI linked to spatial navigation and location processing, encodes semantics such as 'on', 'in', and 'of', as illustrated in Figure 8(c). The OWFA, which is associated with visual word recognition and text processing, is shown in Figure 8(d). However, we did not observe similar results for OWFA, likely because the training data consisted mainly of images of natural scenes, with very few related words or text.

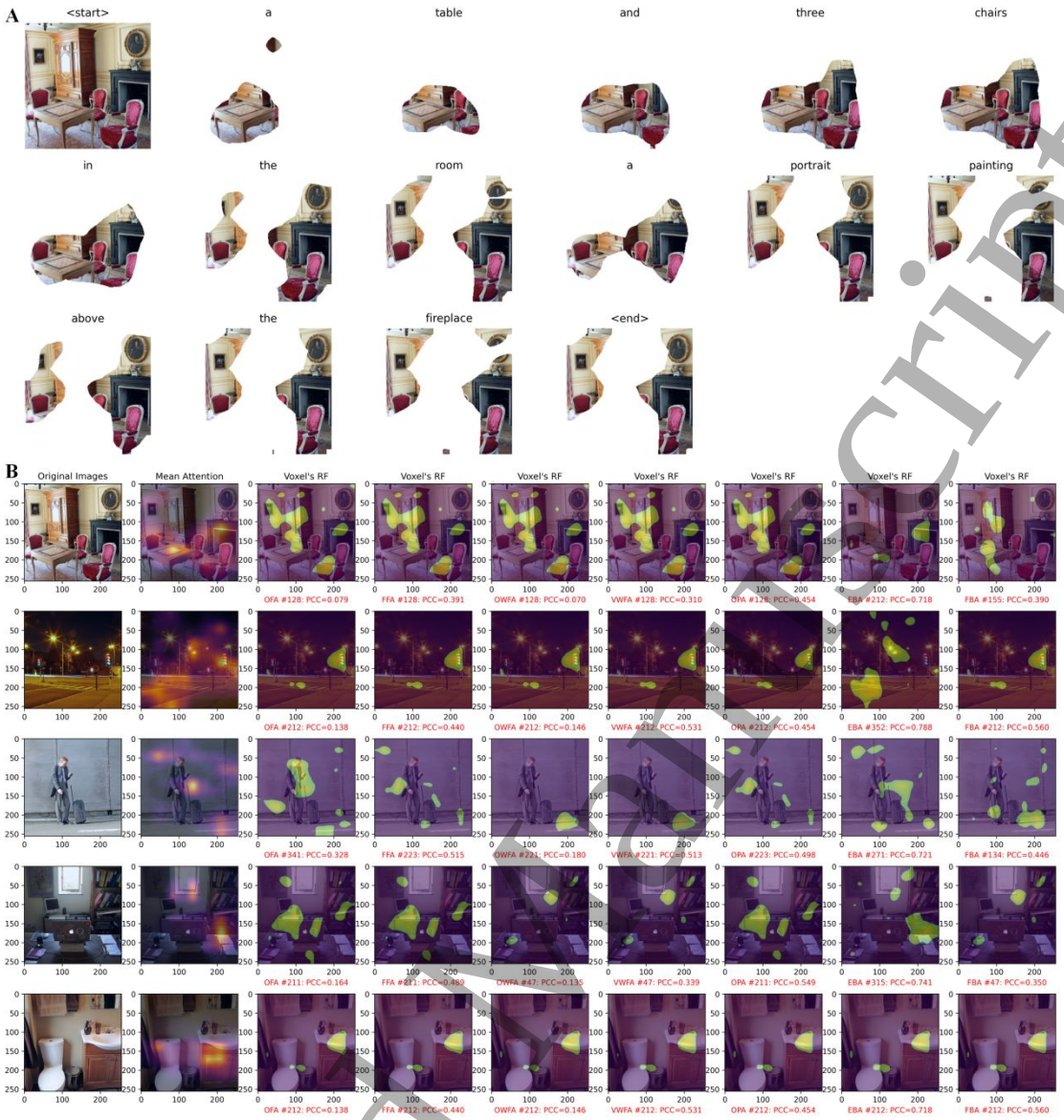

Figure 7. Visualization analysis of attention and RF. (A) Visualization of attention changes in the EBA for subject 1 during the image description process. The 14×14 sparse attention map is resized to 256×256 and overlaid on the original stimulus image. (B) Comparison between the averaged attention map and the RFs of the most predictive voxels across different brain regions for Subject 1. The first column shows the original images, the second column presents the averaged attention maps, and the third to ninth columns display the RF distributions of the most predictive voxels in each brain region.

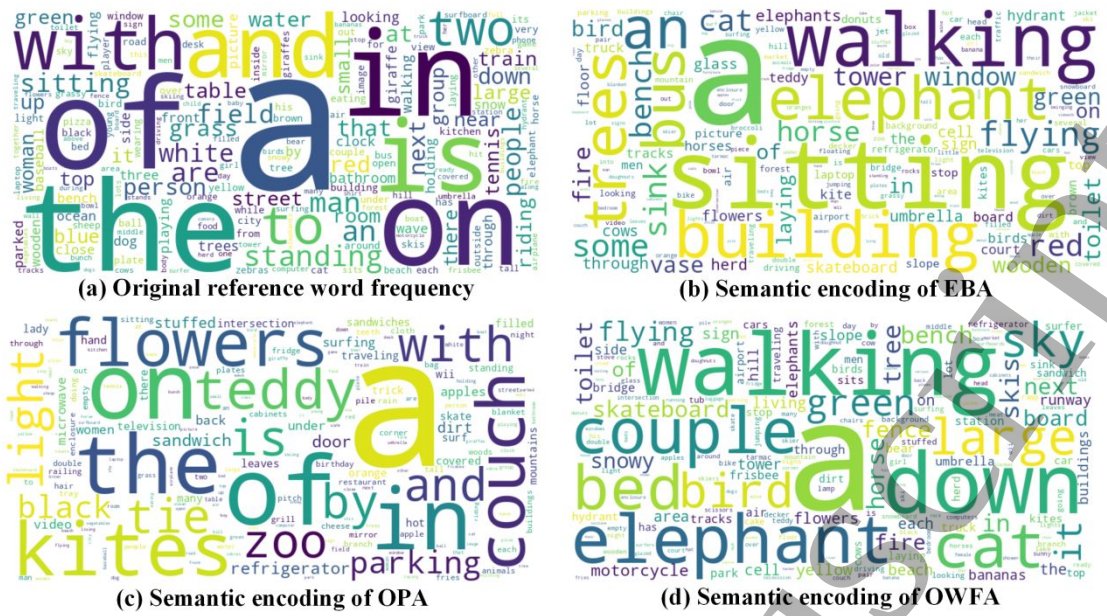

Figure 8. Interpretation of typical ROIs based on related words. Different words are shown in different colors, and the size of each word is proportional to its frequency of occurrence. This visualization reveals the statistical characteristics of words associated with different ROIs. (a) shows the word frequencies of the real texts, while (b), (c), and (d) display the semantic encoding results for EBA, OPA, and OWFA, respectively.

6) Ablation study. The proposed neural network framework comprises three key components: an "Atten" module, which enhances critical semantic regions within ICT; an "RF" module, which models the RF characteristics of each voxel in the neural encoding process; and an "Atten-RF" module, which bridges the representational gap between the two and enables cross-domain information alignment and integration. To evaluate the effectiveness of each component, we conducted an ablation study on the EBA region of Subject 1, who achieved the best overall performance. Specifically, the "w/o Atten + w/ RF + w/o Atten-RF", "w/ Atten + w/o RF + w/o Atten-RF" and "w/ Atten + w/ RF + w/o Atten-RF" refer to simplified versions of the model, where the corresponding modules were removed. In the "w/o RF" configuration, ridge regression is used in place of the RF module to map the feature representations to individual voxel responses, enabling the prediction of brain activity. This method remains one of the most widely adopted and effective strategies in the field of neural encoding. The results of the ablation experiments with different module combinations are presented in Table 3, reported as $mean \pm std$ across five random seeds. The best performance for each metric is highlighted. For statistical analysis, we employed paired two-tailed t-tests to compare the performance differences among models with and without specific modules under the same experimental conditions. The significance threshold was set to $p < 0.05$, where * indicates a significant improvement ($p < 0.05$), ** indicates a highly significant improvement ($p < 0.005$), and – indicates no significant difference ($p \geq 0.05$). All tests were conducted across five random seeds to ensure robustness. The statistical analysis results of different methods are also shown in Table 3. The attention module plays a crucial role in significantly improving the performance of the ICT, which in turn enhances neural encoding to some extent. The RF module design is particularly dominant in this framework. The incorporation of a flexible and learnable RF substantially boosts the neural encoding performance, and surprisingly, it also leads to improvements in image captioning performance. In our opinion, during the neural encoding process, the learnable RF can effectively capture rich semantic features, significantly enhancing the representational power of brain activity. These improved representations offer more accurate and comprehensive semantic information for the ICT when shared across tasks, thereby improving the performance of image captioning. The proposed Atten-RF module further enhances the neural encoding performance without compromising image description accuracy, while also exploring the adaptive adjustment and optimization of the RF through the attention mechanism. In addition, the use of channel attention slightly improves neural encoding performance, though the result is not statistically significant. This may be due to the fact that the feature maps from the low-level visual cortices come from different layers of the neural network, leading to significant feature diversity between channels. In contrast, the feature maps used for neural encoding in the high-level visual cortices in this study are taken from the final feature layer, where the feature diversity between channels is relatively low. As a result, the effect of channel attention is limited.

7) The training of the model. We compared the performance of the feature map extraction module using both frozen and fine-tuned ResNet101, as shown in Figure 9A. It can be observed that the image captioning performance of the fine-tuned model is approximately 4.8% lower than that of the frozen

model in terms of the average Top5-accuracy across all ROIs. However, the neural encoding performance has improved, with the MSE decreasing by about 0.005 and the PCC increasing by about 0.034. The weight parameters of the fine-tuned model are also adjusted to better fit the brain's neural activities, which may compromise some of the captioning generalization capabilities of the ICT. This phenomenon also reflects the trade-off between neural encoding and task optimization.

Taking the EBA of subject 1 as an example, we also present the changes in training loss and performance metrics over different stages of training, as shown in Figure 9B and Figure 9C. It can be observed that the multi-domain joint loss in our model enables fast and stable convergence, with the corresponding Top5-ACC for image description, as well as the MSE and PCC for neural encoding, reaching their peak values.

Table 3. The results of the ablation experiments with different module combinations on the EBA region of Subject 1. The results are shown in the form of the $mean \pm std$ of five random seeds, with the best performance highlighted in bold. A paired two-tailed t-test was employed to compare the performance differences among models with and without specific modules under the same experimental conditions. The significance threshold was set to $p < 0.05$, where * indicates significant improvement ($p < 0.05$), ** indicates highly significant improvement ($p < 0.005$), and – indicates no significant difference ($p \geq 0.05$).

Performance	Neural Encoding			Image Captioning
	MSE↓	PCC↑	R^2↑	Top5-ACC↑
w/o Atten, w/ RF, w/o Atten-RF	.512±.002**	.682±.004**	.413±.002**	.608+.131*
w/ Atten, w/o RF, w/o Atten-RF	.601±.005**	.587±.003**	.352±.004**	.645+.147–
w/ Atten, w/ RF, w/o Atten-RF	.511±.004**	.690±.003**	.418±.007**	.643+.113–
Ours	**.501±.003**	**.696±.002**	**.474±.005**	**.646+.126**

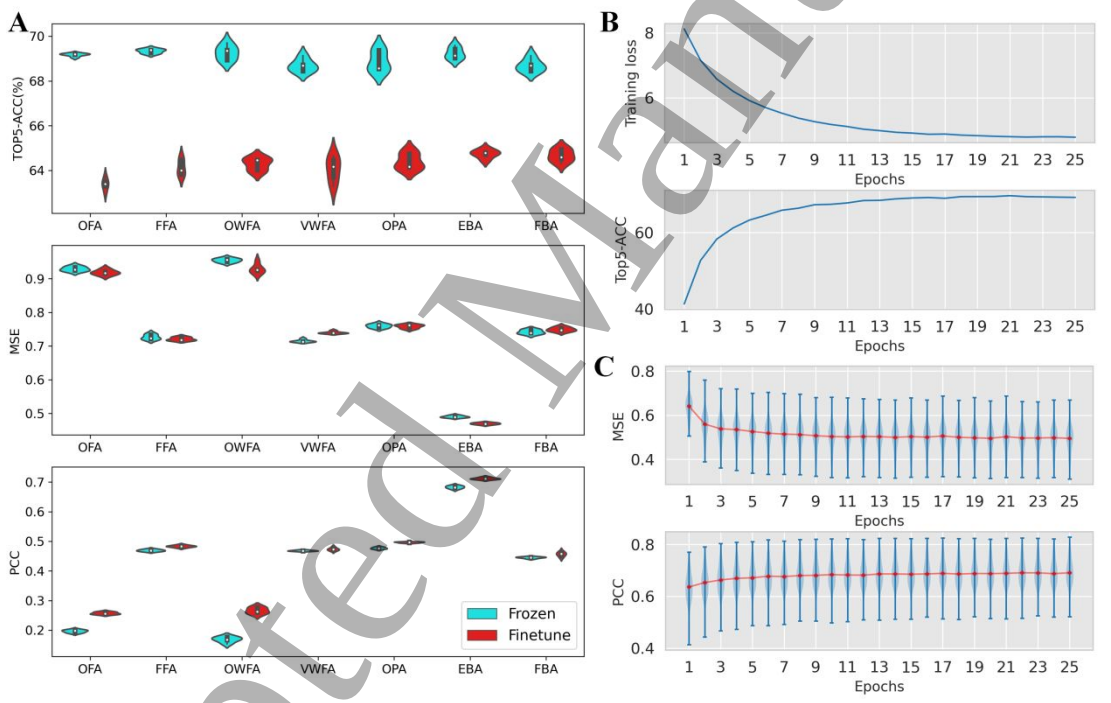

Figure 9. Results of ablation and comparation experiments. (A) Performance comparison of the frozen and fine-tuned feature map extraction modules using ResNet101 for subject 1. (B) During the training process of EBA voxel activity encoding in subject 1, the training loss and Top5-ACC of image captioning performance on the validation set varies with the number of epochs. (C) During the training process of EBA voxel activity encoding in subject 1, the encoding performance (MSE and PCC) of our model on the validation set varies with the number of epochs.

5. Discussion

5.1. Relationship between DNN layers and brain ROIs

Previous neuroscience studies [55, 56] have demonstrated that the ventral and dorsal visual streams are hierarchically organized, with early visual areas responsible for processing low-level features (e.g., edges) and downstream areas encoding increasingly complex features (e.g., shapes). Motivated by this hierarchical organization, we analyzed the contributions of neural network layers at different depths to predicting brain activity across distinct ROIs. We selected three high-level visual regions, namely OPA, EBA, and FFA, as well as four low-level visual regions, including V1, V2, V3, and hV4, all derived from

the ROIs defined in the NSD dataset. This selection established a hierarchical gradient from the early visual cortex to high-level visual areas. V1 through hV4 are located in the occipital cortex and correspond to the early stages of visual processing, where representations progress from basic edge and orientation detection to more complex features such as color and shape. OPA, situated in the occipital lobe, is mainly involved in scene and spatial layout processing and is considered an important transitional region from low-level features to scene semantics. EBA selectively responds to visual information related to bodies and body parts, providing category-specific object representations. FFA, located in the mid-fusiform gyrus, is highly sensitive to facial identity and holistic face representations, reflecting a deeper and more specialized stage of visual processing. By arranging ROIs from low-level to high-level areas, we constructed a biologically grounded hierarchical reference framework. This framework allows us to examine whether features from artificial neural networks of different depths correspond to the stepwise processing in the human visual cortex, and to compare ResNet101 and ResNet-CLIP in terms of hierarchical monotonicity and neuroscientific interpretability. ResNet101, with its explicit layered convolutional structure, provides a clear progression of hierarchical features. In contrast, ResNet-CLIP is trained with a cross-modal alignment objective, and its features exhibit stronger abstraction at the semantic level. For a fair comparison, we selected eight representative layers from both models, covering shallow, intermediate, and deep stages. The selected layers include conv1, layer1-block1 (l1b1), layer1-block3 (l1b3), layer2-block1 (l2b1), layer2-block3 (l2b3), layer3-block1 (l3b1), layer3-block22 (l3b22), and layer4-block3 (l4b3). This choice ensured comprehensive and uniform coverage of the network depth, while aligning with the hierarchical ordering of brain regions from V1 to FFA.

In our analysis, features from different network layers were linearly mapped to voxel responses within each ROI. The PCC between the predicted and actual responses was then calculated, and the attribution method was applied to quantify the relative contribution of each network layer to the prediction accuracy across different ROIs. This approach allowed us to examine the extent to which features at varying depths of artificial networks correspond to the hierarchical representations of the human visual cortex. The layer-wise contributions across network depth are illustrated in Figure 10. The results show that in ResNet101, the contribution in early visual areas (V1–hV4) decreases progressively with increasing layer depth, whereas in high-level visual areas (OPA, EBA, FFA) it increases, forming a clear monotonic hierarchical trend. In contrast, ResNet-CLIP exhibits weaker monotonicity and less pronounced hierarchical gradients. A possible explanation is that the CLIP model incorporates global text-image alignment constraints during training, which disperses the layer-wise features across different visual regions. Although Wang et al. [24] demonstrated that ResNet-CLIP achieves higher encoding performance in high-level visual cortices compared to ResNet101, it lacks the explicit hierarchical structure of ResNet101, which more faithfully reflects the bottom-up hierarchical processing of the visual cortex. This interpretability is crucial for analyzing the neural mechanisms underlying human cognition, and it is also the primary reason why we chose ResNet101 as the backbone network in our study.

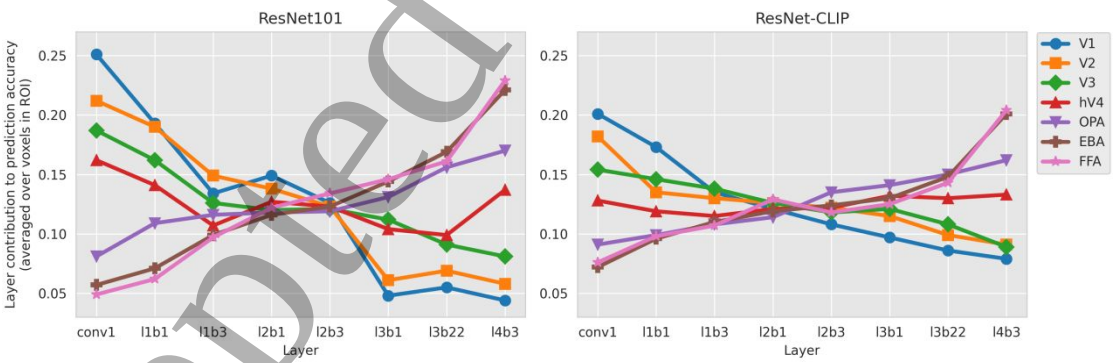

Figure 10. Contributions of different network layers (conv1, l1b1, l1b3, l2b1, l2b3, l3b1, l3b22, and l4b3) in ResNet101 and ResNet-CLIP to predicting voxel responses across ROIs. **Left:** ResNet101 shows a clear monotonic hierarchical trend, with contributions decreasing in early visual areas (V1–hV4) as the layer depth increases, while contributions in high-level visual areas (OPA, EBA, FFA) increase accordingly. **Right:** ResNet-CLIP exhibits weaker monotonicity, with no evident hierarchical gradient.

5.2. Extension for multi-task learning

Multi-Task Learning (MTL) seeks to enhance model performance by simultaneously learning multiple related tasks. MTL has been successfully applied in various fields, including speech recognition, computer vision, and natural language processing. For example, Kokkinos et al. [57] proposed a method that jointly trains multiple visual tasks, including object detection, semantic segmentation, and pose estimation, thereby improving the performance of each individual task. In the field of neural encoding

and decoding, MTL has also been applied and developed. For instance, Huang et al. [58] proposed a novel Visual Language Decoding Model (VLDM) that can simultaneously decode the main categories, semantic labels, and textual descriptions of visual stimuli from visual activity. In this section, we add an image classification task to the original ICT and explore how MTL can improve the model's encoding performance. We obtain the main categories of each stimulus image from the COCO dataset as labels for the image classification task, including 12 categories: "person", "vehicle", "outdoor", "animal", "accessory", "sports", "kitchen", "food", "furniture", "electronic", "appliance", and "indoor". Specifically, we reduce the dimension of the feature maps using global average pooling, followed by two fully connected layers to classify the 12 categories. To facilitate statistical analysis, we selected the OFA and FFA, which exhibited a larger variance in voxel encoding performance in subject 1, for this experiment. The comparisons are shown in Figure 11. In the FFA with better encoding performance, there are 333 significant voxels in the MTL model, in contrast to 331 significant voxels in the original ICT model. In the OFA with poor encoding performance, the number of significant encoding voxels in the MTL model is 188, while that in the original ICT model is 164. Results in both visual brain areas suggest that joint encoding based on MTL may help rescue voxels with poor signal-to-noise ratios. However, the improvement in encoding performance is not significant. This could be due to the addition of a task that is simpler than the ICT. It includes the classification of main categories, the recognition of multiple labels, and the learning of their relationships. More complex image tasks, such as Visual Question Answering (VQA), will be considered in the feature.

This paper combines ICT to improve the neural encoding performance of high-level visual cortices. Strictly speaking, this represents a cross-domain MTL that must account for data heterogeneity and representation alignment. The key distinction between cross-domain MTL and general MTL lies in the complementarity between tasks, cross-domain collaborative modeling, and a design more aligned with biological mechanisms. The experimental results demonstrate that our model exhibits stronger neural encoding ability, higher generalization performance, and deeper explanatory power, marking a significant extension and innovation in the field of MTL.

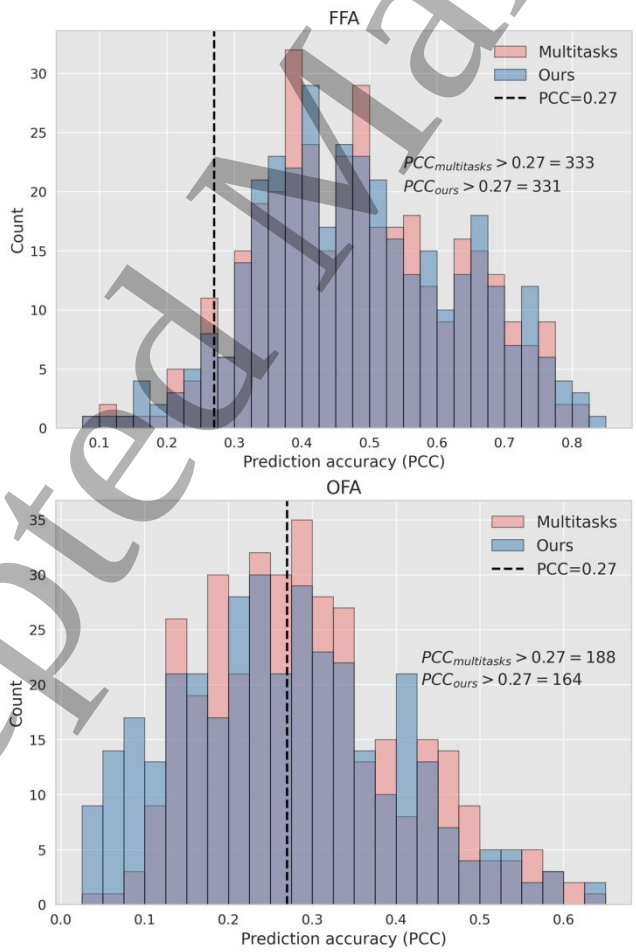

Figure 11. Comparison of encoding performance between the multitask model and our model on the FFA and OFA in subject 1.

5.3. Advantages and disadvantages of our model

There are several possible reasons why our model outperforms the other methods. One key factor is the incorporation of high-level visual tasks, along with the introduction of the attention mechanism, which enhances both performance and interpretability in visual tasks. Another possible reason is the use of flexible RFs. While traditional RFs often rely on strong assumptions, such as a Gaussian distribution, our model enhances flexibility and learnability by making minimal assumptions about the spatial characteristics of RFs. Since high-level visual ROIs process complex and abstract semantic information, our model avoids pre-defining the shape of the RFs. Instead, it leverages the distribution of training data to automatically learn the optimal RF for each voxel, thereby identifying the most relevant spatial locations and ranges for visual features. The third possibility lies in the design of the Attention-RF module, which bridges the gap between two distinct tasks or domains. This not only enhances the model's ability to handle complex tasks, but also enables finer-grained information alignment between images and brain activities at the semantic level.

While our model achieves strong encoding performance, there are still drawbacks worth consideration. The ROI-wise encoding mode greatly improves encoding efficiency by building a unified model to encode all voxels within a given ROI simultaneously. Although we perform preliminary voxel selection during the pre-processing stage, some invalid voxels may still interfere with the encoding process within the same ROI. In the future, the impact of invalid voxels can be further minimized by incorporating more advanced voxel screening methods, such as voxel importance scores based on task relevance or sparsity regularization techniques. In addition, since voxels within the same small-scale brain region tend to perform similar calculations, this suggests the potential to incorporate Graph Neural Networks (GNN) to model the spatial and functional correlations between voxels. This multi-voxel-wise encoding approach may allow for the retention of voxels with poor signal-to-noise ratios, without sacrificing the model's computational efficiency.

6. Conclusion

DNNs trained on image classification tasks have achieved success in neural encoding studies of the early visual cortex. However, their performance in high-level visual areas remains suboptimal due to the complexity and abstract nature of semantic features encoding in these regions. To address this challenge, we propose a novel end-to-end neural encoding model based on Image Captioning Tasks (ICT) to enhance the encoding performance of the high-level visual cortex. Our method incorporates an attention module to focus on key image pixes and an RF module to model voxel-specific visual fields. Additionally, we introduce the Atten-RF module to bridge the domain gap between visual stimuli and brain responses, facilitating joint optimization of the visual and neural components. Experimental results demonstrate that the proposed model outperforms existing state-of-the-art approaches, achieving superior neural encoding performance in high-level visual areas. Furthermore, visualization analyses of RF distributions and semantic encoding characteristics highlight the biological interpretability of our approach.

However, it is important to note that the ICT in this study was trained exclusively on visual stimuli from the NSD dataset, which may constrain its generalizability to broader contexts. Future work could consider incorporating large language models such as GPT-4 [59] and LLaMA [60]. These models provide rich multimodal representations that may help extract more fine-grained perceptual features. By combining them with limited brain activity data, it may be possible to further enhance neural encoding performance and improve the robustness and generalization of brain–machine models.

CRediT authorship contribution statement

Xu Yin: Software, Methodology, Writing – original draft. **Jiuchuan Jiang:** Conceptualization, Methodology. **Sheng Ge:** Methodology, Writing – review & editing. **John Q. Gan:** Validation, Writing – review & editing. **Haixian Wang:** Writing – review & editing, Formal analysis, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

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/Neural-encoding-for-high-level-visual-cortices).

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 92270113 and 62176054.

References

[1] Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. NeuroImage 2011;56(2):400-10.
[2] Poldrack RA, Farah MJ. Progress and challenges in probing the human brain. Nature 2015;526(7573):371-9.
[3] Qian D, Zeng H, Cheng W, Liu Y, Bikki T, Pan J. NeuroDM: Decoding and visualizing human brain activity with EEG-guided diffusion model. Comput Methods Programs Biomed 2024;251:108213.
[4] Li D, Wei C, Li S, Zou J, Qin H, Liu Q. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In: Proc. Annual Conference on Neural Information Processing Systems (NeurIPS); 2024.
[5] Stokes MG, Wolff MJ, Spaak E. Decoding rich spatial information with high temporal resolution. Trends Cogn Sci 2015;19(11):636-8.
[6] Benchetrit Y, Banville H, King JR. Brain decoding: Toward real-time reconstruction of visual perception. In: Proc. International Conference on Learning Representations (ICLR); 2024.
[7] Zhang YJ, Yu ZF, Liu JK, Huang TJ. Neural decoding of visual information across different neural recording modalities and approaches. Mach Intell Res 2022;19(5):350-65.
[8] Li R, Li J, Wang C, Liu H, Liu T, Wang X, et al. Multi-semantic decoding of visual perception with graph neural networks. Int J Neural Syst 2024;34(4):2450016.
[9] Ferrante M, Boccato T, Passamonti L, Toschi N. Retrieving and reconstructing conceptually similar images from fMRI with latent diffusion models and a neuro-inspired brain decoding model. J Neural Eng 2024;21(4):046001.
[10] Luo J, Cui W, Liu J, Li Y, Guo Y, Xu S. Visual image decoding of brain activities using a dual attention hierarchical latent generative network with multiscale feature fusion. IEEE Trans Cogn Dev Syst 2023;15(2):761-73.
[11] St-Yves G, Allen EJ, Wu Y, Kay K, Naselaris T. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. Nat Commun 2023;14(1):3329.
[12] Wen H, Shi J, Chen W, Liu Z. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. Sci Rep 2018;8(1):3752.
[13] Shi J, Wen H, Zhang Y, Han K, Liu Z. Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision. Hum Brain Mapp 2018;39(5):2269-82.
[14] Ahonen T, Hadid A, Pietikainen M. Face description with local binary patterns: Application to face recognition. IEEE Trans Pattern Anal Mach Intell 2006;28(12):2037-41.
[15] Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 2011; 21(19):1641-6.
[16] Huth AG, Nishimoto S, Vu AT, Gallant JL. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 2012;76(6):1210-24.
[17] Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep 2016;6:27755.
[18] Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. Nat Neurosci 2019;22:1761-70.
[19] Himberger KD, Chien HY, Honey CJ. Principles of temporal processing across the cortical hierarchy. Neurosci 2018;389:161-74.
[20] St-Yves G, Naselaris T. The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. NeuroImage 2018;180:188-202.
[21] Seeliger K, Ambrogioni L, Güçlütürk Y, Bulk LM, Güçlü U, Gerven MJ. End-to-end neural system identification with neural information flow. PLoS Comput Biol 2021;17(2):e1008558.
[22] Wang C, Yan H, Huang W, Sheng W, Wang Y, Fan YS, et al. Neural encoding with unsupervised spiking convolutional neural network. Commun Biol 2023;6(1):880.
[23] Ma S, Wang L, Hou S, Zhang C, Yan B. Large-scale parameters framework with large convolutional kernel for encoding visual fMRI activity information. Cereb Cortex 2024;34(7):bhae257.
[24] Wang AY, Kay K, Naselaris T, Michael JT, Leila W. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. Nat Mach Intell 2023;5:1415-26.
[25] Li J, Zhang C, Wang L, Ding P, Hu L, Yan B, et al. A visual encoding model based on contrastive self-supervised learning for human brain activity along the ventral visual stream. Brain Sci 2021;11(8):1004.
[26] Wang H, Huang L, Du C, Li D, Wang B, He H. Neural encoding for human visual cortex with deep neural networks learning 'What' and 'Where'. IEEE Trans Cogn Dev Syst 2021;13(4):827-40.
[27] Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y. BabyTalk: Understanding and generating simple image descriptions. IEEE Trans Pattern Anal Mach Intell 2013;35(12):2891-903.
[28] Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017.
[29] Seeliger K, Ambrogioni L, Güçlütürk Y, Bulk LM, Güçlü U, Gerven MAJ. End-to-end neural system identification with neural information flow. PLoS Comput Biol 2021; 17(2):e1008558.

[30] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16×16 words: Transformers for image recognition at scale. In: Proc. International Conference on Learning Representations (ICLR); 2021.
[31] Al-Shamayleh AS, Adwan O, Alsharaiah MA, Hussein AH, Kharma QM, Eke CI. A comprehensive literature review on image captioning methods and metrics based on deep learning technique. Multimed Tools Appl 2024;83:34219-68.
[32] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015.
[33] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. Show, attend and tell: Neural image caption generation with visual attention. in: Proc. International Conference on Machine Learning (ICML); 2016.
[34] Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom-up and top-down attention for image captioning and visual question answering. in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018.
[35] Yao T, Pan Y, Li Y, Mei T. Exploring visual relationship for image captioning. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018.
[36] Ma S, Wang L, Chen P, Qin R, Hou L, Yan B. A mixed visual encoding model based on the larger-scale receptive field for human brain activity. Brain Sci 2022;12(12):1633.
[37] Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature 2008;452(7185):352-5.
[38] Xue M, Wu X, Li J, Li X, Yang G. A convolutional neural network interpretable framework for human ventral visual pathway representation. In: Proc. AAAI Conference on Artificial Intelligence; 2024.
[39] Qiao K, Zhang C, Chen J, Wang L, Tong L, Yan B. Effective and efficient ROI-wise visual encoding using an end-to-end CNN regression model and selective optimization. In: Proc. Conference on Human Brain and Artificial Intelligence (HBAI); 2021.
[40] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015.
[41] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 1998;20(11):1254-9.
[42] Corbetta M, Shulman GL. Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 2002;3(3):201-15.
[43] Klink PC, Chen X, Vanduffel W, Roelfsema PR. Population receptive fields in nonhuman primates from whole-brain fMRI and large-scale neurophysiology in visual cortex. eLife 2021;10:e67304.
[44] Kriegeskorte N. Deep neural networks: A new framework for modeling biological vision and brain information processing. Annu Rev Vis Sci 2015;1:417-46.
[45] Du C, Du C, Huang L, He H. Reconstructing perceived images from human brain activities with Bayesian deep multiview learning. IEEE Trans Neural Netw Learn Syst 2019;30(8):2310-23.
[46] Wu H, Zhu Z, Wang J, Zheng N, Chen B. An encoding framework with brain inner state for natural image identification. IEEE Trans Cogn Dev Syst 2021;13(3):453-64.
[47] Wu H, Zheng N, Chen B. Feature-Specific Denoising of Neural Activity for Natural Image Identification. IEEE Trans Cogn Dev Syst 2022;14(2):629-38.
[48] Allen EJ, St-Yves G, Wu Y, Breedlove JL, Prince JS, Dowdle LT, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nat Neurosci 2022;25(1):116-26.
[49] Lu Y, Du C, Wang D, He H. MindDiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In: Proc. ACM International Conference on Multimedia (ACMMM); 2023.
[50] Du C, Fu K, Li J, He H. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. IEEE Trans Pattern Anal Mach Intell 2023;45(9):10760-77.
[51] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39(4):640-51.
[52] Gao JS, Huth AG, Lescroart MD, Gallant JL. Pycortex: An interactive surface visualizer for fMRI. Front Neuroinform 2015;9:23-35.
[53] Cohen MR, Maunsell JHR. Using neuronal populations to study the mechanisms underlying spatial and feature attention. Neuron 2011;70(6):1192-204.
[54] Darcet T, Oquab M, Mairal J, Bojanowski P. Vision transformers need registers. In: Proc. International Conference on Learning Representations (ICLR); 2024.
[55] Güçlü U, Gerven MA. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci 2015;35(27):10005-14.
[56] Güçlü U, Gerven MA. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. Neuroimage 2017;145:329-36.
[57] Kokkinos I. Ubernet: training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017.
[58] Huang W, Yang P, Tang Y, Qin F, Li H, Wu D, et al. From sight to insight: A multi-task approach with the visual language decoding model. Inf Fusion 2024;112:102573.
[59] Zhu D, Chen J, Shen X, Li X, Elhoseiny M. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. 2023;arXiv preprint arXiv:2304.10592.
[60] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M, Lacroix T, et al. LLaMA: Open and efficient foundation language models. 2023;arXiv preprint arXiv:2302.13971.