Received: 20 April 2021 | Revised: 24 June 2021 | Accepted: 13 July 2021

DOI: 10.1002/hbm.25603

RESEARCH ARTICLE

WILEY

A dual-channel language decoding from brain activity with progressive transfer training

Wei Huang[1] | Hongmei Yan[1] | Kaiwen Cheng[2] | Yuting Wang[1] | Chong Wang[1] | Jiyi Li[1] | Chen Li[3] | Chaorong Li[1] | Zhentao Zuo[4] | Huafu Chen[1]

[1]The Clinical Hospital of Chengdu Brain Science Institute, MOE Key Lab for Neuroinformation, High-Field Magnetic Resonance Brain Imaging Key Laboratory of Sichuan Province, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China

[2]School of Language Intelligence, Sichuan International Studies University, Chongqing, China

[3]Department of Medical Information Engineering, Sichuan University, Chengdu, China

[4]State Key Laboratory of Brain and Cognitive Science, Beijing MR Center for Brain Research, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China

Correspondence
Hongmei Yan and Huafu Chen, The Clinical Hospital of Chengdu Brain Science Institute, MOE Key Lab for Neuroinformation, High-Field Magnetic Resonance Brain Imaging Key Laboratory of Sichuan Province, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China.
Email: hmyan@uestc.edu.cn (H. Y.) and chenhf@uestc.edu.cn (H. C.)

Zhentao Zuo, State Key Laboratory of Brain and Cognitive Science, Beijing MR Center for Brain Research, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China.
Email: ztzuo@bcslab.ibp.ac.cn (Z. Z.)

Funding information
Strategic Priority Research Program of Chinese Academy of Science, Grant/Award Number: XDB32010300; Ministry of Science and Technology of China, Grant/Award Numbers: 2019YFA0707103, 2020AAA0105601;

Abstract

When we view a scene, the visual cortex extracts and processes visual information in the scene through various kinds of neural activities. Previous studies have decoded the neural activity into single/multiple semantic category tags which can caption the scene to some extent. However, these tags are isolated words with no grammatical structure, insufficiently conveying what the scene contains. It is well-known that textual language (sentences/phrases) is superior to single word in disclosing the meaning of images as well as reflecting people's real understanding of the images. Here, based on artificial intelligence technologies, we attempted to build a dual-channel language decoding model (DC-LDM) to decode the neural activities evoked by images into language (phrases or short sentences). The DC-LDM consisted of five modules, namely, Image-Extractor, Image-Encoder, Nerve-Extractor, Nerve-Encoder, and Language-Decoder. In addition, we employed a strategy of progressive transfer to train the DC-LDM for improving the performance of language decoding. The results showed that the texts decoded by DC-LDM could describe natural image stimuli accurately and vividly. We adopted six indexes to quantitatively evaluate the difference between the decoded texts and the annotated texts of corresponding visual images, and found that Word2vec-Cosine similarity (WCS) was the best indicator to reflect the similarity between the decoded and the annotated texts. In addition, among different visual cortices, we found that the text decoded by the higher visual cortex was more consistent with the description of the natural image than the lower one. Our decoding model may provide enlightenment in language-based brain-computer interface explorations.

KEYWORDS
artificial intelligence, functional magnetic resonance imaging, language decoding, progressive transfer, visual cortex

National Natural Science Foundation of China, Grant/Award Numbers: 31730039, U1808204, 62036003, 61773094; Key Project of Research and Development of Ministry of Science and Technology, Grant/Award Number: 2018AAA0100705

1 | INTRODUCTION

With the development of brain imaging technologies, such as Electroencephalography (EEG), the local field potential (LFP), Magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI), neural decoding is becoming a hot topic in brain-computer interfaces (BCIs) (Ahmadi, Davoudi, Behroozi, & Daliri, 2020; Dezfouli & Daliri, 2020). Visual system is an important part of the human brain for understanding natural world. More than half of the human brain's cortex is involved in visual cognition and information processing. The visual information obtained by the brain affects a series of cognitive activities such as emotion, social communication, and decision-making (Bernardini, Porayska-Pomsta, & Smith, 2014; Gasper, 2004; Jack & Schyns, 2017). Understanding the mechanism of human visual cognition and information processing as well as decoding the visual perceptual activities has always been one of the most cutting-edge and challenging research directions in brain science (Cox & Savoy, 2003; Huang et al., 2018).

At present, the field of visual decoding mainly focuses on extracting single/multiple semantic category tags contained in a scene from visual neural activity, namely single-tag decoding and multiple-tag decoding.

The purpose of single-tag decoding is to obtain the main category tag in the scene from visual neural activities (Behroozi & Daliri, 2014; Horikawa, Tamaki, Miyawaki, & Kamitani, 2013; Jafakesh, Jahromy, & Daliri, 2016; Taghizadeh-Sarabi, Daliri, & Niksirat, 2015). Haxby et al. used functional magnetic resonance imaging (fMRI) to record visual nerve activities by presenting subjects with different types of images, and found that the neural activity varied with different types of image (Haxby et al., 2001). Therefore, neuroscientists hypothesized that visual nerve activity measured by fMRI contained rich information and could be used to extract category tags of different natural scenes. Subsequently, a large number of researches have studied the category decoding of natural image stimuli by fMRI (Cox & Savoy, 2003; Huang, Yan, Wang, Li, Yang, et al., 2021; Qiao et al., 2019; Song, Zhan, Long, Zhang, & Yao, 2011). For example, Cox et al. used a support vector machine (SVM) to classify the visual nerve activity induced by images and achieved a good accuracy of decoding (Cox & Savoy, 2003). Immediately afterwards, other researchers further discovered that neural activity induced by other visual stimuli, such as stripe orientation (Kamitani & Tong, 2005) and spatial frequency (Kamitani & Tong, 2005), could also be decoded into the categories of stimuli. Even dreams can also been decoded into single-tag based visual imagery categories (Horikawa et al., 2013). Single-tag decoding usually only extracts the main category tags in visual scenes from the visual activity.

Multiple-tag decoding is an upgraded version of single-tag decoding, which is designed to extract not only the main category in the scene, but also other semantic categories, such as object, action, color or size (Fukuma et al., 2018; Hu, Guo, Han, & Liu, 2015; Huth et al., 2016; Nishida & Nishimoto, 2018; Stansbury, Naselaris, & Gallant, 2013). In early days, Gallant team conducted an exploratory study on fMRI-based multiple-tag decoding with dynamic movies as visual stimuli (Huth, Nishimoto, Vu, & Gallant, 2012). They systematically drew the semantic representation maps of 1,705 objects and action categories in the cerebral cortex, and then used the regression model to achieve multiple-tag decoding of the dynamic video. More recently, Wang et al. compared the effects of different stimuli (words, sentences, and images) on the decoding performance of multiple-tag semantic category, and found that the brain activity induced by images was more accurate than that induced by words and sentences (Wang, Zhang, Wang, Lin, & Zong, 2020). However, although multiple-tag decoding can disclose more semantic information in visual scenes from neural activities, it cannot specify the relationship between those decoded category tags, neither semantically nor grammatically.

In brief, the semantic category tags obtained by the above two decoding methods are isolated and grammarless, which cannot fully reflect the mutual relationships among tags in natural images. By comparison, textual language (sentences/phrases) has advantages in revealing the meanings of images and the relationships between objects and actions accurately and vividly. What's more, language can better reflect people's real understanding for the images. Therefore, in this article, language decoding is carried out to decode the neural activities evoked by images into phrases or short sentences with grammatical structure and semantic coherence, which contains richer information and clearer relationship between different semantic category tags. We assume that the fMRI neural activity is a special kind of human language, through which we can realize the conversion from neural activities to textual language based on relevant technologies of machine translation (Bahdanau, Cho, & Bengio, 2015; Cho et al., 2014; Makin, Moses, & Chang, 2020; Papineni, Roukos, Ward, & Zhu, 2002). Here, we developed a Dual Channel Language Decoding Model (DC-LDM) based on Transformer to translate visual neural activity induced by natural images into text. In addition, we adopted a progressive transfer strategy to train DC-LDM to improve the performance of language decoding. Finally, we investigated the performance of different visual cortices on language decoding.

2 | EXPERIMENTS AND METHODS

2.1 | Dataset

In order to achieve fMRI-based language decoding, we followed the framework shown in Figure 1. We first used fMRI to record brain

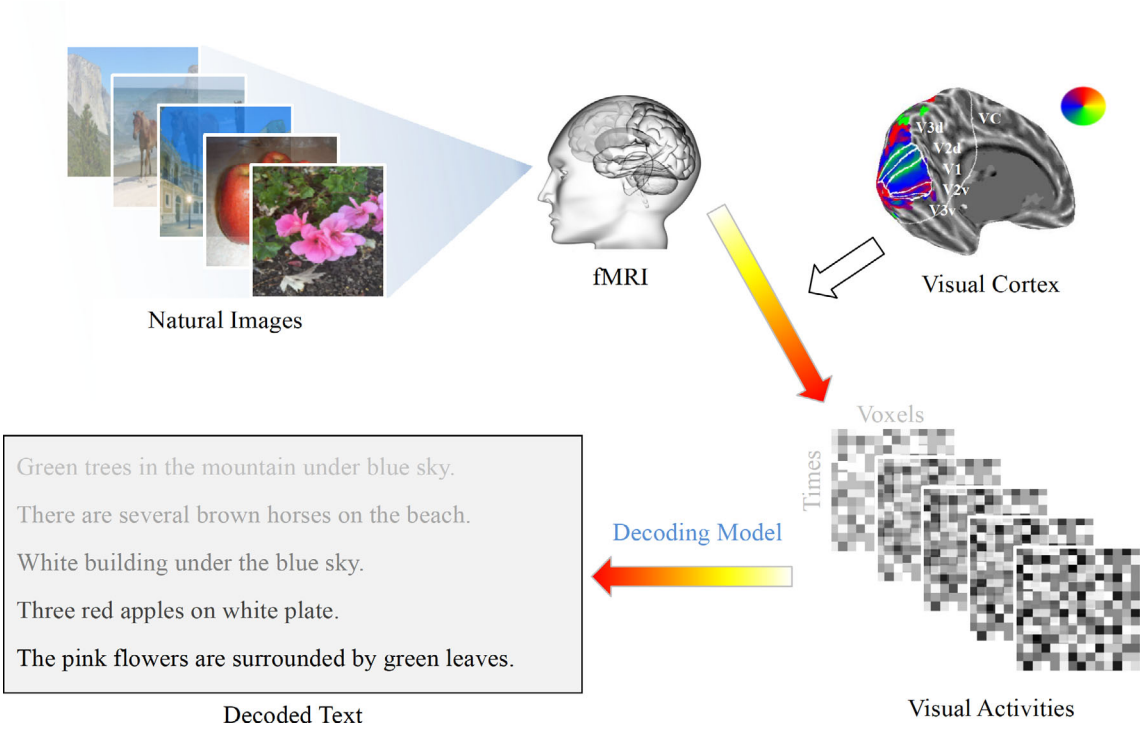

FIGURE 1 An overview of the language decoding. It consists of three parts: (1) recording the brain activity evoked by natural images through fMRI; (2) extracting the neural activity of the visual cortex; (3) decoding the text from the visual activity through a decoding model fMRI, functional magnetic resonance imaging

activities while participants were viewing natural images. The retinotopic experiment was also done to locate the spatial position of the visual area in the brain (Dumoulin & Wandell, 2008; Huang, Yan, Wang, Li, Yang, et al., 2020). Then, for each image, we took out the multi-time response patterns of the corresponding visual cortex. Finally, the multi-time visual response patterns were fed into our language decoding model to generate the textual language. The experimental dataset consisted of 2,750 natural images with a resolution of 256 × 256 pixels, which were selected from ImageNet (Deng et al., 2009). The corresponding fMRI data contains visual neural activities from the V1, V2, V3, LVC, HVC, and VC areas of five subjects (3 males and 2 females; average age 25). VC denotes the entire occipital cortex. LVC denotes the combination of V1, V2, and V3. HVC denotes the remainder of VC minus LVC. For the convenience of calculation, the number of voxels in each visual area was unified to 2,000. For visual areas (V1, V2, and V3) with fewer than 2,000 voxels, 2,000 voxels are interpolated by up-sampling. For visual areas (LVC, HVC, and VC) with greater than 2,000 voxels, 2,000 voxels are selected by the F-score feature selection algorithm (Huang et al., 2018; Polat & Güneş, 2009). In addition, the 14 s visual activities were measured by fMRI after the appearance of the image stimulation. Therefore, the neural activity of each visual area is a matrix with dimensions of 14 × 2,000 (time × voxels). All subjects were provided written informed consent before the MRI experiments, and protocols were approved by the Institutional Review Board of the Institute of Biophysics, Chinese Academy of Sciences. A detailed experimental description is shown in our previous work (Huang, Yan, Wang, Li, Yang, et al., 2020; Huang, Yan, Wang, Li, Zuo, et al., 2020; Huang, Yan, Wang, Yang, et al., 2021; C. Wang et al., 2020).

Besides, five annotators (3 males and 2 females; average age 22) gave a textual description for each image, which was used as a reference for later model training and testing. In short, our dataset contained 2,750 samples, of which each had 1 natural image, 5 texts (from five annotators), and 6 matrices (from six visual areas). The dataset was randomly divided into 2,500 training and 250 test samples. Note that for each sample, we only randomly selected one of the five texts from the annotators to participate in the training of the model.

2.2 | Function definition

Before building our language decoding model (refer to previous research [Vaswani et al., 2017]), we first define four functions, namely "Multi-Head Attention Function," "Multi-Head Attention for Decoding Function," "Feed-Forward Network Function," and "Positional Encoding Function."

2.2.1 | Multi-head attention function ($\mathbb{F}_{MHA}$)

Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. We

assume $X_{in} \in \mathbb{R}^{T \times d_{model}}$ denotes the input. The output of $\mathbb{F}_{MHA}$ is as follows:

$$\mathbb{F}_{MHA}(X_{in}) = \text{Concat}(H_1,...,H_h)W^O$$

$$H_i = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right)V_i$$

$$Q_i, K_i, V_i = X_{in}W_i^Q, X_{in}W_i^K, X_{in}W_i^V$$

where W_i^Q, W_i^K, W_i^V, and W^O are parameter matrices, $W_i^Q \in \mathbb{R}^{d_{model} \times d_q}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$. d_{model} represents the number of rows in W_i^Q, W_i^K, and W_i^V. d_q, d_k, and d_v represent the number of columns in W_i^Q, W_i^K, and W_i^V, respectively. h denotes the number of heads.

2.2.2 | Multi-head attention for decoding function ($\mathbb{F}_{MHAD}$)

We assume $X_{in1} \in \mathbb{R}^{T \times d_{model}}$ and $X_{in2} \in \mathbb{R}^{T \times d_{model}}$ denote two inputs. The output of $\mathbb{F}_{MHAD}$ is as follows:

$$\mathbb{F}_{MHAD}(X_{in1}, X_{in2}) = \text{Concat}(H_1,...,H_h)W^O$$

$$H_i = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right)V_i$$

$$Q_i, K_i, V_i = X_{in1}W_i^Q, X_{in2}W_i^K, X_{in2}W_i^V$$

similarly, where $W_i^Q \in \mathbb{R}^{d_{model} \times d_q}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ are parameter matrices. d_{model} represents the number of rows in three parameter matrices (W_i^Q, W_i^K, and W_i^V). d_q, d_k, and d_v represent the number of columns in W_i^Q, W_i^K, and W_i^V, respectively. h denotes the number of heads.

2.2.3 | Feed-forward networks function ($\mathbb{F}_{FFN}$)

We assume $X_{in} \in \mathbb{R}^{T \times d_{model}}$ denotes the input. The output of $\mathbb{F}_{FFN}$ is as follows:

$$\mathbb{F}_{FFN}(X_{in}) = \max(0, X_{in}W_1 + b_1)W_2 + b_2$$

where W_1 and W_2 are parameter matrices. b_1 and b_2 are parameter vectors.

2.2.4 | Positional encoding function (PE)

In order to make use of the sequence order based on the transformer model, Vaswani et al. proposed the positional encoding (Vaswani et al., 2017) as follows:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10{,}000^{2i/d_{model}}}\right)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10{,}000^{2i/d_{model}}}\right)$$

where sin and cos denote the sine and cosine functions, respectively. pos denotes the position of the token in the sequence.

In this study, the number of heads (h) is 8. To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce outputs of dimension $d_{model} = 512$. For each of these we use $d_q = d_k = d_v = d_{model}/h = 64$. The values of the above hyperparameters are based on previous research on Transformers (Vaswani et al., 2017).

2.3 | The dual-channel language decoding model

Here, we propose a Dual-Channel Language Decoding Model (DC-LDM) in Figure 2, which contains five modules, namely "Image-Extractor," "Image-Encoder," "Nerve-Extractor," "Nerve-Encoder," and "Language-Decoder." The first channel (image-channel), including "Image-Extractor" and "Image-Encoder," aims to extract the semantic features of natural images ($I \in \mathbb{R}^{L \times W \times C}$). L, W, and C denote the length, width, and number of channels of the image respectively. The second channel (nerve-channel), including "Nerve-Extractor" and "Nerve-Encoder," aims to extract the semantic features of visual activities ($X = [x_1,...,x_T]^{\mathrm{T}} \in \mathbb{R}^{T \times M}$). T and M denote the time length and the number of voxels of visual activities, respectively. In the training phase, the corresponding outputs of the two channels are weighted by the transfer factor (α) to "Language-Decoder."

2.3.1 | Image/Nerve Extractor

The "Image-Extractor" is a CNN model which aims to extract the features of natural images. The input and output of the CNN are the natural image (I) and corresponding image-features ($L^{(Img)} \in \mathbb{R}^{T \times M}$), respectively. To reduce the parameters of the CNN model, the natural image is downsampled to $32 \times 32 \times 3$. For the CNN, leaky ReLU with leakiness 0.2 is adopted in all layers, except for the last layer where linear activation is used. Batch normalization is conducted in the network, except for the last layer. T is set as the durations of multi-time visual response patterns. In this study, for each image being processed, the continuous 14 s visual response patterns measured by the fMRI are collected. Therefore, T equals 14.

The "Nerve-Extractor" is a BiGRU which aims to extract the features of brain activities. The input and output of the BiGRU are visual activities (X) and corresponding neural feature ($L^{(Neu)} \in \mathbb{R}^{T \times M}$), respectively.

2.3.2 | Image/Nerve Encoder

The "Image-Encoder" contains two sub-layers (multi-head attention and fully connected feed forward network). The operations of residual

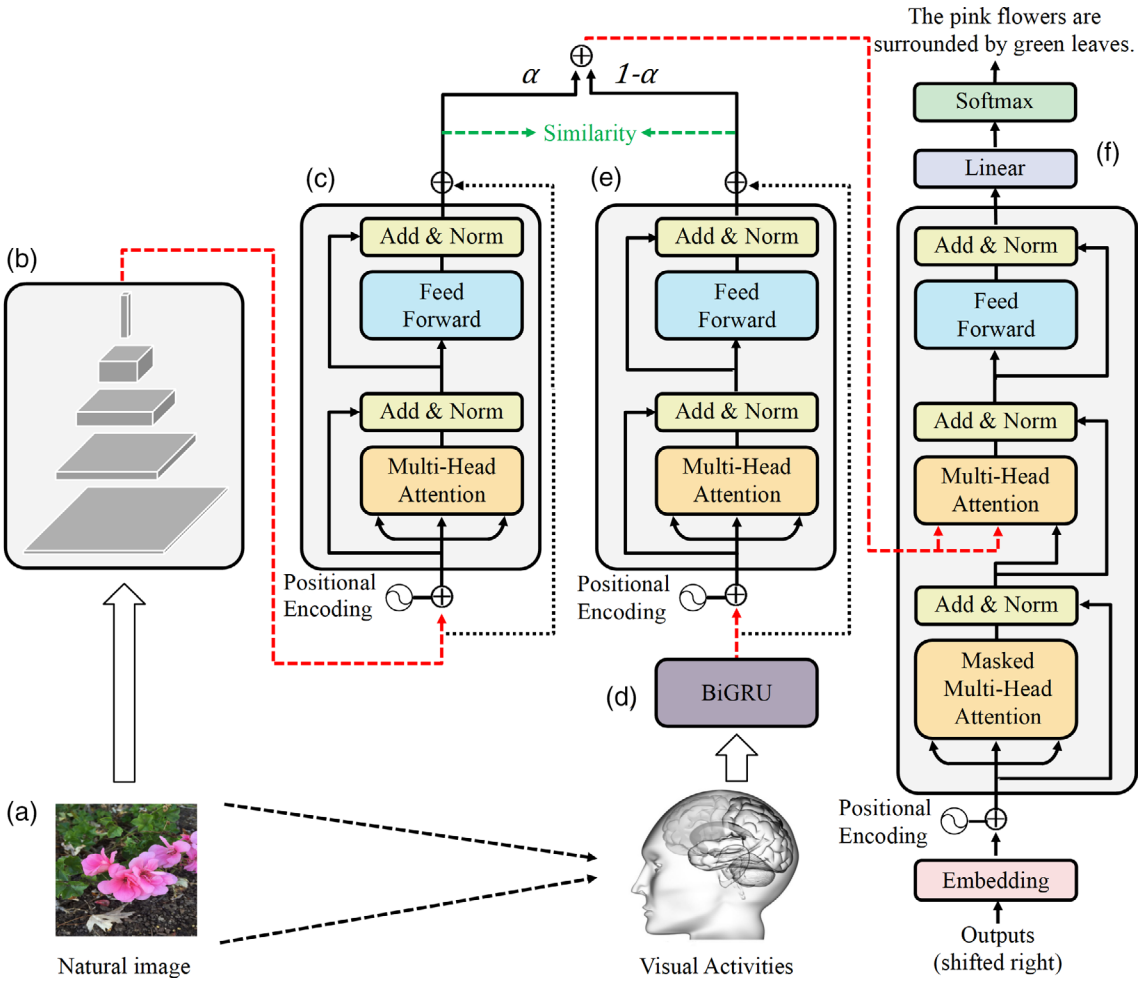

FIGURE 2 The structure of the dual-channel language decoding model. (a) Natural image and its corresponding visual activities (multi-time and multi-voxel) measured by fMRI. (b) The "Image-Extractor" which is built with a CNN model. (c) "Image-Encoder" which aims to further extract the latent features of natural images. (d) The "Nerve-Extractor" which is built with a Bidirectional Gate Recurrent Unit (BiGRU) model. (e) The "Nerve-Encoder" which aims to further extract the latent features of visual activities measured by fMRI. (f) The "Language-Decoder" which decodes the weighted features (latent features of natural image and visual activities) into text. fMRI, functional magnetic resonance imaging

connection (He, Zhang, Ren, & Sun, 2016) and layer normalization (Ba, Kiros, & Hinton, 2016) are used after the two sub-layers. The initial input of "Image-Encoder" is $H_0^{(\mathrm{Img})} = L^{(\mathrm{Img})} + \mathrm{PE}\left(L^{(\mathrm{Img})}\right)$. The outputs ($C_1^{(\mathrm{Img})}$ and $H_1^{(\mathrm{Img})}$) of two sub-layers are sequentially calculated as:

$$C_1^{(\mathrm{Img})} = \mathrm{LN}\left(\mathbb{F}_{\mathrm{MHA}}\left(H_0^{(\mathrm{Img})}\right) + H_0^{(\mathrm{Img})}\right)$$

$$H_1^{(\mathrm{Img})} = \mathrm{LN}\left(\mathbb{F}_{\mathrm{FFN}}\left(C_1^{(\mathrm{Img})}\right) + C_1^{(\mathrm{Img})}\right)$$

where LN and PE denote the layer normalization and positional encoding. The final output of "Image-Encoder" is $H_1^{(\mathrm{Img})} + L^{(\mathrm{Img})}$.

"Nerve-Encoder" has the same network structure as "Image-Encoder." The initial input of "Nerve-Encoder" is $H_0^{(\mathrm{Neu})} = X + \mathrm{PE}(X)$. Refer to the description above, the final output of "Nerve-Encoder" is $H_1^{(\mathrm{Neu})} + L^{(\mathrm{Neu})}$.

2.3.3 | Language-Decoder

Compared with "Image-Encoder", "Language-Decoder" adds a new sub-layer. The outputs ($C_1^{(\mathrm{Dec})}$, $D_1^{(\mathrm{Dec})}$ and $H_1^{(\mathrm{Dec})}$) of three sub-layers are sequentially calculated as:

$$C_1^{(\mathrm{Dec})} = \mathrm{LN}\left(\mathbb{F}_{\mathrm{MHA}}\left(H_0^{(\mathrm{Dec})}\right) + H_0^{(\mathrm{Dec})}\right)$$

$$b1 = \alpha \times \left(H_1^{(\mathrm{Img})} + L^{(\mathrm{Img})}\right)$$

$$b2 = (1-\alpha) \times \left(H_1^{(\mathrm{Neu})} + L^{(\mathrm{Neu})}\right)$$

$$D_1^{(\mathrm{Dec})} = \mathrm{LN}\left(\mathbb{F}_{\mathrm{MHAD}}\left(C_1^{(\mathrm{Dec})}, b1 + b2\right) + C_1^{(\mathrm{Dec})}\right)$$

$$H_1^{(\mathrm{Dec})} = \mathrm{LN}\left(\mathbb{F}_{\mathrm{FFN}}\left(D_1^{(\mathrm{Dec})}\right) + D_1^{(\mathrm{Dec})}\right)$$

where α denote the transfer factor, which dynamically changes with each epoch. Finally, a fully connected network with the Softmax is supposed to predict the generated text (Y), as follows:

$$Y = \text{Softmax}\left(H_1^{(\text{Dec})} W + b\right)$$

where W and b are parameter matrix and parameter vector, respectively.

2.4 | Objective function

Objective function is extremely important for deep learning, because it guides the learning and representation of network parameters by back-propagation of the error between the predicted results and the real markers. One of the highlights of this article is the definition of loss function. Here, the objective function includes a text prediction loss and a similarity loss. The text prediction loss is used to help our model better generate text that represents visual perception. The similarity loss aims to make the latent features of the image consistent with the latent features of visual activities as well as possible. We use the Sparse-Softmax Cross-Entropy (SSCE) and the Cosine Similarity (CSIM) as the text prediction loss and the similarity loss for the training. The objective function Loss is defined as follows:

$$\text{Loss} = \text{SSCE}\left(Y, \widehat{Y}\right) - \text{CSIM}\left(H_1^{(\text{Img})}, H_1^{(\text{Neu})}\right)$$

where Y and $\widehat{Y}$ denote the real text and the predicted text, respectively. $H_1^{(\text{Img})}$ and $H_1^{(\text{Neu})}$ denote the final outputs of "Image-Encoder" and "Nerve-Encoder", respectively.

2.5 | Training strategy of progressive transfer

In addition to the network structure of the language decoding model, good training strategies may improve the decoding performance greatly. The training strategy of our DC-LDM consists of two stages, namely a transfer stage and a decoding stage. There are a total of 400 epochs in the training process. In the transfer stage (1st–200th epochs) in Figure 3a, the transfer factor (α) gradually decreases from 0.5 to 0. By adjusting α, the weight of the image-channel is gradually transferred to that of the nerve-channel until it disappears. In the decoding stage (201th–400th epochs) in Figure 3b, α remains 0. During the test in Figure 3c, only visual activities are input to DC-LDM to generate decoded text. In addition, the gradient descent algorithm is used to optimize our model (learning rate: 0.08; batch size: 64).

Here, the proposed DC-LDM is implemented in TensorFlow. The size of the model is about 278 M. The training time is about 1 hr using one computer with one Nvidia TITAN X(Pascal) GPU with 12 GB Ram.

3 | RESULTS

3.1 | Decoded texts

Through the iterative process of 400 epochs, DC-LDM gradually converges. The loss curve (also known as the convergence curve) from one subject as an example during training is showed in Figure S1. Then, we fix the parameters of the DC-LDM and use the test set to get the decoded result. Figure 4 shows the decoding results of different images in the test set with VC fMRI activities from a sample subject. It can be seen that the decoded texts by our model are reasonable for describing the natural images, although they are not completely consistent with annotator's texts. The results show that our proposed model can capture the semantic information from visual activities and represent it through textual language.

3.2 | Quantitative evaluation

In order to find the most suitable index for evaluating the language decoding model, we adopted six indexes to evaluate the difference between the decoded texts and the annotator's texts. The six indicators are as follows:

1. BLEU (bilingual evaluation understudy [Papineni et al., 2002]) is an indicator that is used to evaluate the difference between the sentences generated by the model and the actual sentences. It is often used in the evaluation of machine translation tasks in natural language processing.
2. ROUGE (Recall-Oriented Understudy for Gisting Evaluation [Lin, 2004]) is an evaluation method for automatic summarization. Its calculation process is similar to BLEU.
3. CIDEr (consensus-based image description evaluation [Vedantam, Lawrence Zitnick, & Parikh, 2015]) is an evaluation commonly used for image caption generation.
4. WCS (Word2vec-Cosine similarity). This method first averages all word vectors from Word2vec (Mikolov, Chen, Corrado, & Dean, 2013) in the sentence to obtain the sentence representation. Then, the cosine similarity algorithm is used to calculate the similarity between the two sentence representations.
5. GCS (Glove-Cosine similarity). This method first averages all word vectors from Glove (Pennington, Socher, & Manning, 2014) in the sentence to obtain the sentence representation. Then, the cosine similarity algorithm is used to calculate the similarity between the two sentence representations.
6. FTCS (FastText-Cosine similarity). This method first averages all word vectors in the sentence from FastText (Joulin, Grave, Bojanowski, & Mikolov, 2016) to obtain the sentence representation. Then, the cosine similarity algorithm is used to calculate the similarity between the two sentence representations.

Here, we used the above evaluation indicators to calculate two vectors respectively: (a) A decoding-vector (1,250 dimensions)

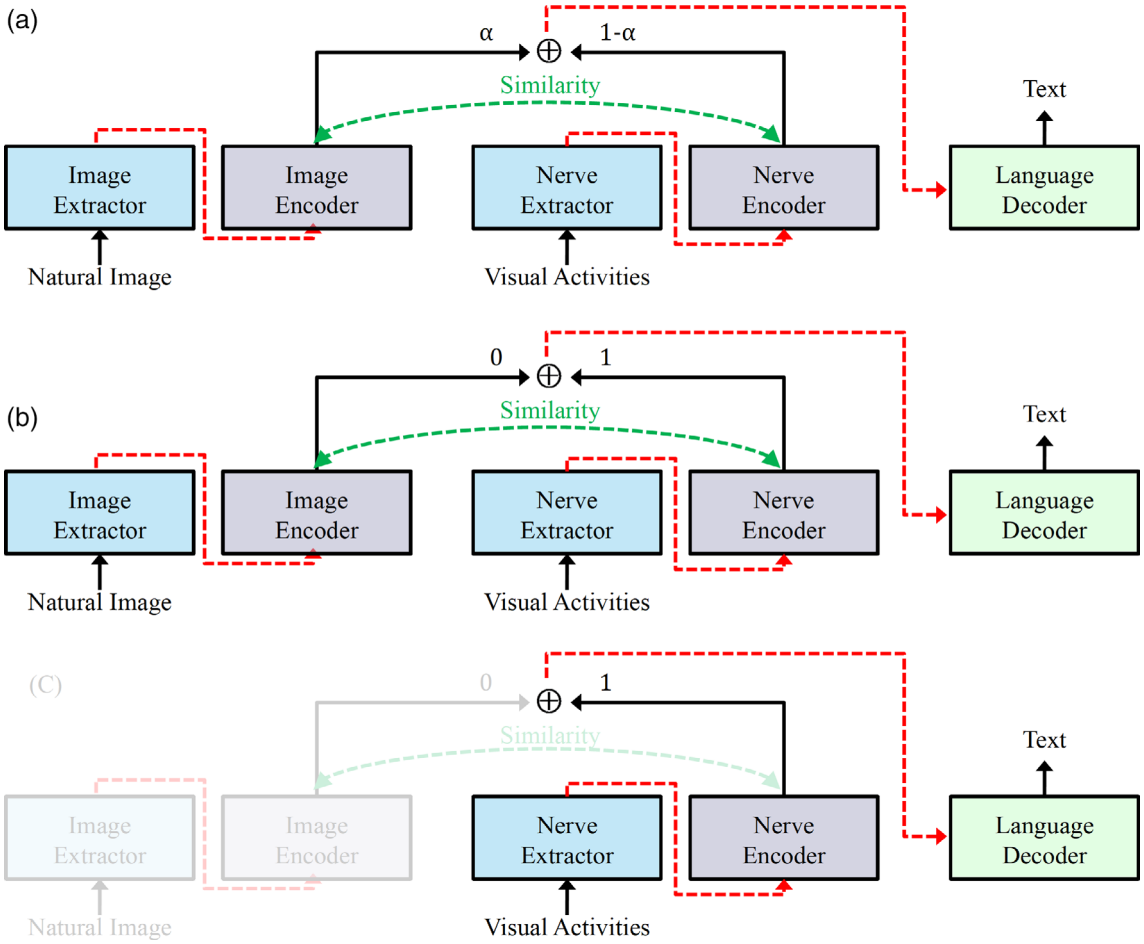

FIGURE 3 The training strategy of the dual-channel language decoding model. (a) The transfer stage for training. (b) The decoding stage for training. (c) Language model testing stage

generated by the decoded texts (size: 250) and the corresponding annotator's texts (size: 5) together; (b) A baseline-vector (3,125,000 dimensions) generated by the decoded texts (size: 250) and the training's texts (2,500 × 5). The two vectors were obtained by the VC activities from each subject. Theoretically, the bigger difference between the decoding-vector and the baseline-vector is, the better decoding performance is. The frequency histograms from the two vectors were calculated. Figure 5 shows the frequency histograms corresponding to the decoding-vector and the baseline-vector with the six evaluation indicators from subject 1 as an example. Intuitively, the results show that the frequency histograms of the two vectors obtained by WCS have the greatest difference, which indicates that WCS may be the most suitable indicator to evaluate the performance of the language decoding. Similar results from the other four subjects were shown in the Figures S2–S5.

For a quantitative comparison, we obtained the normalized decoding-vectors by subtracting the mean of the baseline vectors and then divide them by the variance of the baseline vector. The average value of the normalized decoding-vectors obtained by the six evaluation indicators from the five subjects are shown in Figure 6. The results show that WCS is still the most suitable indicator for evaluating the performance of language decoding.

3.3 | Function of similarity loss

We introduced similarity loss in the objective loss. The similarity loss was obtained by calculating the cosine similarity between the outputs of the "Image-Encoder" and the "Nerve-Encoder." In order to verify its effect, we compared the decoding performance with or without similarity loss. In Figure 7, the results show that the decoding performance with similarity loss is significantly higher (paired T test, $p < .05$) than those without.

3.4 | Comparison of training strategies

In the language decoding model, we first weighted the output of "Image-Encoder" and "Nerve-Encoder," and their corresponding weights were α and $1-\alpha$ respectively. The weighted item was then used as input into a "language decoder" to generate the text description of the image stimulus. The training strategy of our decoding model was special. α changes dynamically with each epoch ($0.5 \rightarrow 0$). To illustrate the effect of this particular training strategy, we compared it with two other training strategies: (a) Training strategy I, α is maintained at 0.5 during the training; (b) Training strategy II, α

1.Yellow flower with light green leaves.
2.An orange flower.
3.An orange flower.
4.An orange flower with leaves.
5.Small flower of an orange petal.

■ A beautiful and elegant flower.

1.Horses playing on the yellow grassland.
2.There are two brown horses in the sun.
3.Two wild horses on the yellow grass.
4.Two cuddly brown horses on a beautiful prairie.
5.Two brown horses on a yellow and green meadow.

■ A group of brown horses stand on the lush grass.

1.Sail boat on the sea water under the yellow sunset.
2.There is an orange sunset glow in the blue sky , and some white ships on the blue sea water.
3.Boat floating on the water under the blue sky.
4.Several speed boats on the water.
5.A deep blue sea water and several brown boats.

■ A group of dark boats standing on the sea water.

1.Lawn and building under blue sky.
2.There is a white european building on the green grass.
3.Green lawn and building.
4.White building with a green lawn under blue sky.
5.A green lawn and european buildings.

■ White building with beautiful lawn.

1.white flowers on green branches with green leaves.
2.There is a white flower on the branch.
3.There is a white flower among the branches covered with snow.
4.Blooming flower on a branch.
5.There are several small white flowers in front of a clump of green bushes.

■ A small white flower placed between the green leaves.

1.Red apple against white background.
2.A red apple.
3.A round, crystal red apple.
4.A red apple.
5.There is a big, red apple in front of the whiteboard.

■ A round red apple.

1.Stacked oranges together.
3.Some orange oranges.
3.A bunch of fresh and seductive oranges.
4.A bunch of fresh oranges.
5.There are a bunch of orange oranges in front of the brown wooden board.

■ A bunch of fresh oranges.

1.Buildings with white pavement under blue sky.
2.There are many european buildings on the white road.
3.Empty white street under blue sky and buildings next to it.
4.Building with a clock under the sun.
5.There are a pile of beige buildings on the gray ground under the blue sky.

■ Buildings and pavement under the blue sky.

1.River water with rocks and dim sky.
2.Silver lake water and black stones under the silver sky.
3.Smooth stones in the water and grey sky.
4.Dark stones in the water under grey sky.
5.A gray and black sky , sea water and stones.

■ Water surface and stones in the sky.

1.Zebra standing on the yellow grass beside the tree.
2.Green trees in front of a zebra.
3.Zebra standing on the grass.
4.A zebra is standing in front of the green tree.
5.There is a zebra on the white grass.

■ A cute zebra in the woods.

1.Small white flowers on a branch.
2.There are three white flowers on the brown land.
3.The green shoots are covered with white flowers.
4.A bunch of white flowers.
5.Three white flowers and a few gray stones.

■ A small white flower on a brown branch.

1.Green trees on the mountain under blue sky.
2.There is a brown wilderness in the mountain under the blue sky.
3.Red coniferous mountain forest under the blue sky.
4.Mountain grasslands and woods under the blue sky.
5.There is a green and yellow mountain forest under the blue sky.

■ The mountains and lush forests under the blue sky.

FIGURE 4 Text descriptions of natural image stimuli. For each image, the first five texts (black) are from the five annotators. The text at the bottom (blue) is decoded from a subject's VC activities by our decoding model

is maintained at 0 during the training. We referred to the training process of progressive transfer in this article as Strategy III. Figure 8 shows the comparison of decoding performance when using the three training strategies. The result shows that the decoding performance of Strategy III is significantly higher (paired T test, $p < .05$) than that of the other two strategies.

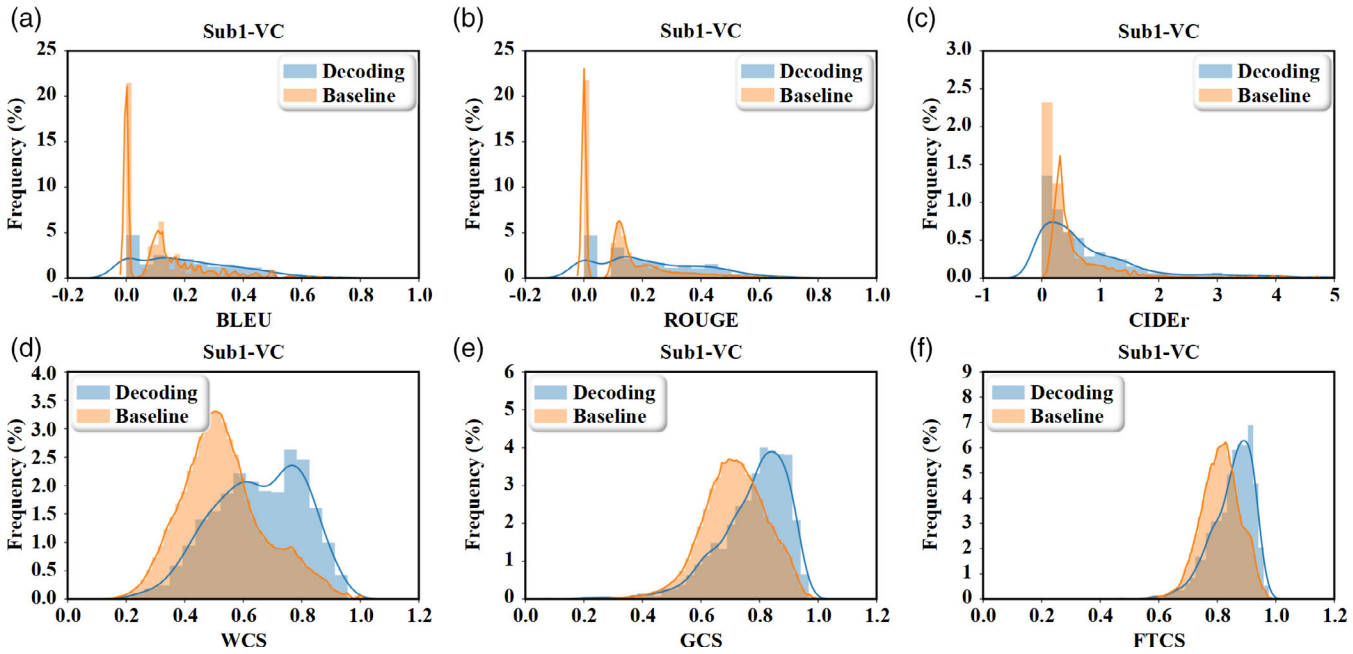

FIGURE 5 Frequency histogram of different evaluation indicators from Subject-1's VC activities. Panels (a), (b), (c), (d), (e), and (f) are the frequency histograms obtained using BLEU, ROUGE, CIDEr, WCS, GCS, and FTCS respectively. In each subgraph, the blue distribution represents the frequency histogram of the evaluation value between the decoded text and the corresponding annotator's text from test set. The orange distribution, also called the baseline distribution, represents the frequency histogram of the evaluation value between the decoded text and all the text from the training set. BLEU, bilingual evaluation understudy; CIDEr, consensus-based image description evaluation; FTCS, FastText-Cosine similarity; GCS, Glove-Cosine similarity; ROUGE, recall-oriented understudy for gisting evaluation; WCS, Word2vec-Cosine similarity

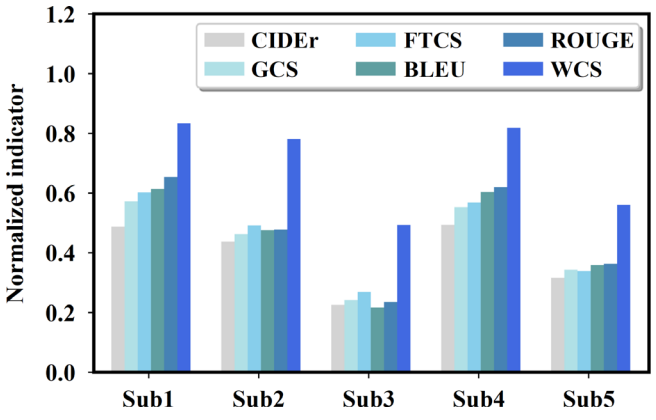

FIGURE 6 Comparison of different evaluation indexes. Different bars represent the mean values of normalized decoding-vectors obtained by different evaluation indicators

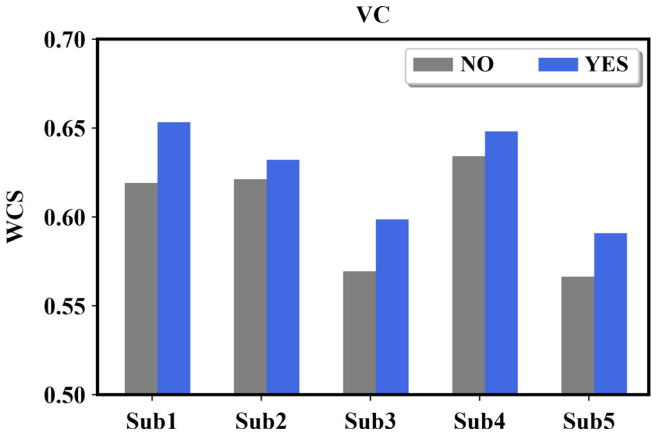

FIGURE 7 Decoding performance of VC activities from five subjects under different loss functions. The gray bar and blue bar represent the average WCS obtained without and with similarity loss function, respectively. WCS, Word2vec-Cosine similarity

3.5 | Comparison of decoding performance of different visual areas

Finally, to investigate the function of different visual areas in language decoding, we also compared the decoding performance of different visual areas signals. We first used neural activities from V1, V2, V3, LVC, HVC, and VC to obtain the decoded texts respectively. Then, WCS was used as the indicator of the decoding-vector and baseline-vector corresponding to each visual area. The frequency histograms of the two vectors obtained by different visual areas from Subject 1 are shown in Figure 9. Intuitively, the result shows that the frequency histograms of the two vectors from HVC and VC have the largest difference. The greater difference between the two frequency histograms indicates the better decoding performance of the visual area. Similar results of the other four subjects are shown in Figures S6–S9. For quantitative comparison, we also show the average WCS of different visual areas from five subjects in Figure 10. The results show that the average WCS of HVC and VC is higher than that of V1, V2, V3, and LVC.

4 | DISCUSSION

In this article, we designed a novel language decoding model that could decode visual activity in the brain as measured by fMRI into text. We selected six evaluation indicators to evaluate the performance of language decoding. The results showed that the text decoded by DC-LDM from fMRI activity could describe the natural image accurately and vividly. Through the comparison of frequency histograms, we found that WCS was the best indicator to reflect the similarity between the decoded text and the annotated text corresponding to the visual image.

The objective function is very important for deep learning models, and its significance lies in the guidance provided for the optimization of model parameters. In DC-LDM, we obtained the latent features of natural images and visual activities through image-channel (contains "Image-Extractor" and "Image-Encoder") and nerve-channel (contains "Nerve-Extractor," "Nerve-Encoder"), respectively. The neural system has its own understanding and processing for natural images, so the semantic information contained in the recorded visual activities is an indirect representation of image stimuli. To achieve the best results, the semantic information of visual activities should be consistent with that of natural images. In the actual model training process, it is necessary to calculate the similarity between two latent features and define it into the objective function. Moreover, the gradient descent method made the two latent features gradually become similar and achieved our goal of pursuing semantic information consistency. We compared the decoding performance with or without similarity loss in the total loss function through ablation analysis, and the result also proved that the inclusion of similarity loss improved the performance of language decoding. The definition of loss function like this could effectively extract the common semantic information for decoding from the stimulus and the corresponding neural activity, and might provide a reference for other decoding models.

In addition, we explored the impact of different training strategies on decoding performance. We found that the training strategy of progressive transfer performed best. The progressive transfer training strategy started with the same weight of the image channel and

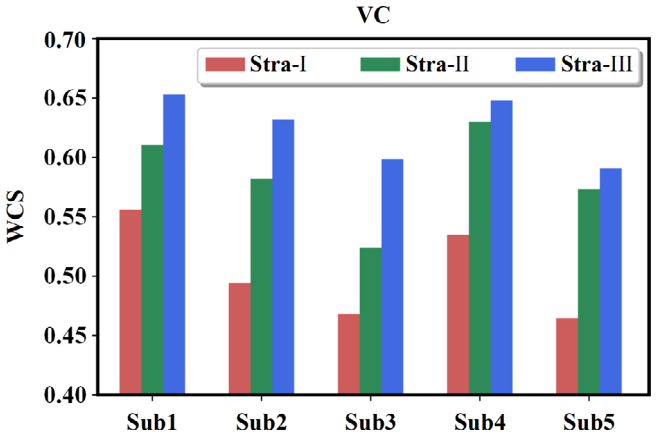

FIGURE 8 Decoding performance of different training strategies using VC activities from five subjects. The red, green, and blue bars represent the average WCS obtained by Strategy I, Strategy II, and Strategy III, respectively. WCS, Word2vec-Cosine similarity

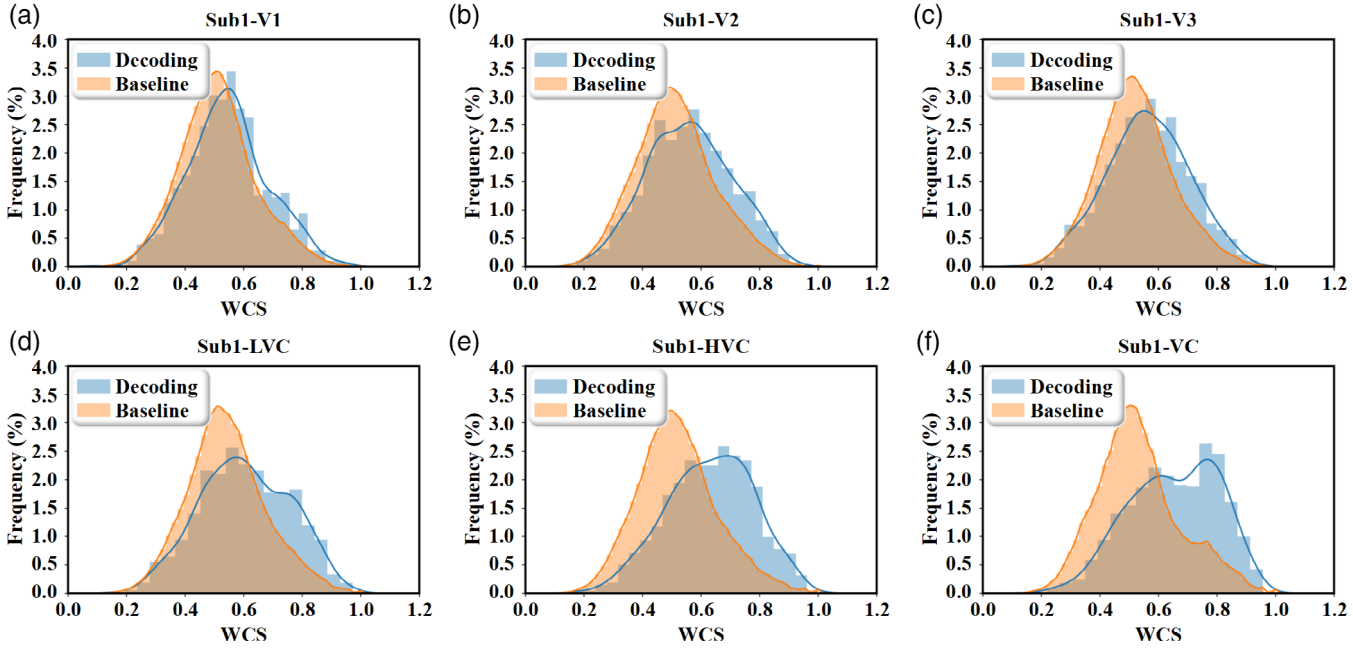

FIGURE 9 Frequency histogram of different visual area from Subject-1. Panels (a), (b), (c), (d), (e), and (f) are the frequency histograms obtained using V1, V2, V3, LVC, HVC, and VC respectively. In each subgraph, the blue distribution represents the frequency histogram of the WCS value between the decoded text and the corresponding annotator's text from the test set. The orange distribution represents the frequency histogram of the WCS value between the decoded text from the test set and all the text from the training set

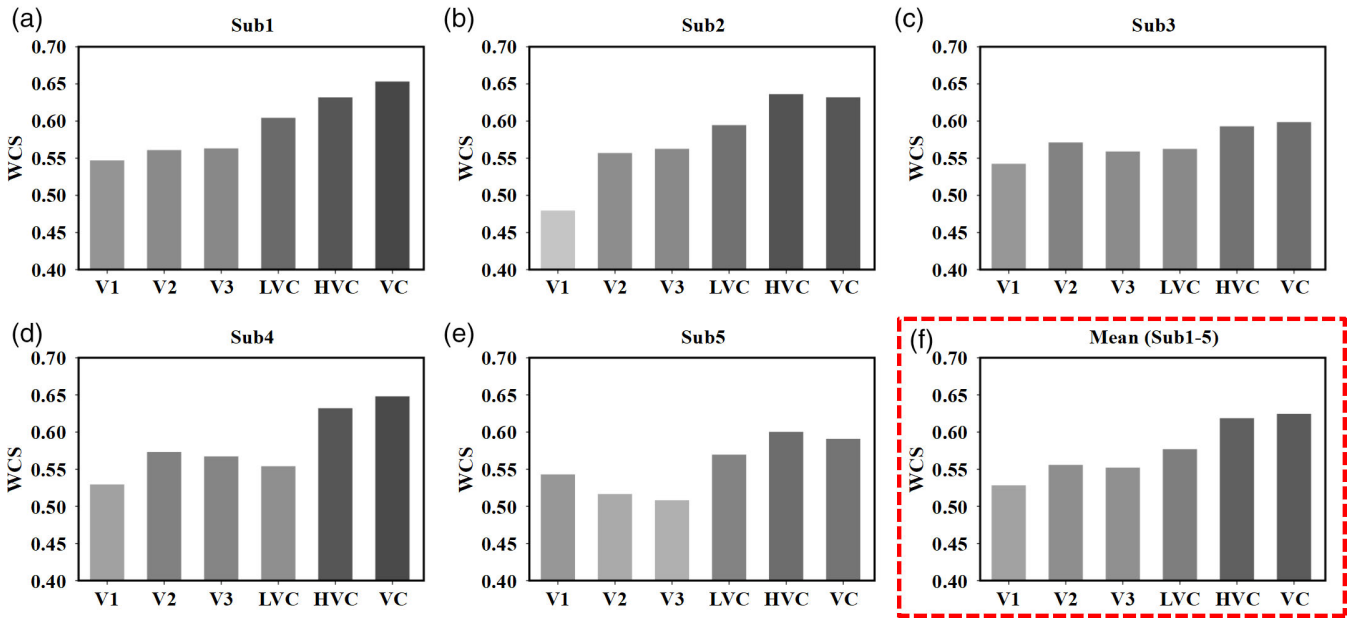

FIGURE 10 Decoding performance of different visual area. Panels (a), (b), (c), (d), and (e), respectively, represent the average WCS of V1, V2, V3, V4, LVC, HVC, and VC from five subjects. Panel (f) represents the average decoding performance of five subjects in different visual areas. WCS, Word2vec-Cosine similarity

nerve-channel, and then smoothly transferred the weight from the image-channel to the nerve-channel, as shown in Figure 3a. The essence of this progressive transfer was to gradually guide the training of the model of neural activities to the text by training the model of the image to the text (image captioning [Cho, Courville, & Bengio, 2015; Vinyals, Toshev, Bengio, & Erhan, 2015]). The idea of progressive transfer comes from a progressive growing GAN (Karras, Aila, Laine, & Lehtinen, 2017), which uses a progressive growth training strategy to start with low resolution images, and then gradually increase the resolution by adding layers to the network to generate high-resolution images. All in all, the training strategy of progressive transfer may provide important reference for the decoding model. In particular, for the multi-channel decoding model, when the weight of one channel needs to be transferred to another channel, the training strategy of progressive transfer may be crucial.

Finally, by comparing different visual areas, we found that the decoding performance of the high-level visual cortex (HVC and VC) is significantly higher than that of low-level visual areas (V1, V2, V3, and LVC). It is once again confirmed that the high-level visual cortex contains more semantic information than the low-level visual cortex (Huang, Yan, Wang, Li, Yang, et al., 2020), even when this semantic information is decoded into text.

ACKNOWLEDGMENTS

This work was supported in part by Key Project of Research and Development of Ministry of Science and Technology (2018AAA0100705), the National Natural Science Foundation of China (61773094, 62036003, U1808204, and 31730039), the Ministry of Science and Technology of China (2020AAA0105601 and 2019YFA0707103), and the Strategic Priority Research Program of Chinese Academy of Science (XDB32010300).

CONFLICT OF INTEREST

The authors declare no competing financial interests.

DATA AVAILABILITY STATEMENT

We will release the data at the appropriate time.

ORCID

Huafu Chen https://orcid.org/0000-0002-4062-4753

REFERENCES

Ahmadi, A., Davoudi, S., Behroozi, M., & Daliri, M. R. (2020). Decoding covert visual attention based on phase transfer entropy. *Physiology & Behavior*, *222*, 112932.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv Preprint*, 1607.06450. https://arxiv.org/pdf/1607.06450.pdf.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *arXiv Preprint*, 1409.0473. https://arxiv.org/pdf/1409.0473v6.pdf.

Behroozi, M., & Daliri, M. R. (2014). Predicting brain states associated with object categories from fMRI data. *Journal of Integrative Neuroscience*, *13*(4), 1–23.

Bernardini, S., Porayska-Pomsta, K., & Smith, T. J. (2014). ECHOES: An intelligent serious game for fostering social communication in children with autism. *Information Sciences*, *264*, 41–60.

Cho, K., Courville, A., & Bengio, Y. (2015). Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, *17*(11), 1875–1886.

Cho, K., Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv Preprint*. https://arxiv.org/pdf/1406.1078.pdf.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*(2), 261–270.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Dezfouli, M. P., & Daliri, M. R. (2020). Single-trial decoding from local field potential using bag of word representation. *Brain Topography*, *33*(1), 10–21.

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage*, *39*(2), 647–660.

Fukuma, R., Yanagisawa, T., Nishimoto, S., Tanaka, M., Yamamoto, S., Oshino, S., … Kishima, H. (2018). Decoding visual stimulus in semantic space from electrocorticography signals. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Toronto, Ontario*.

Gasper, K. (2004). Do you see what I see? Affect and visual information processing. *Cognition and Emotion*, *18*(3), 405–421.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://10.1109/CVPR.2016.90

Horikawa, T., Tamaki, M., Miyawaki, Y., & Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science*, *340*(6132), 639–642.

Hu, X., Guo, L., Han, J., & Liu, T. (2015). Decoding semantics categorization during natural viewing of video streams. *Proceedings of IEEE Transactions on Autonomous Mental Development*, *7*(3), 201–210.

Huang, W., Yan, H., Liu, R., Zhu, L., Zhang, H., & Chen, H. (2018). F-score feature selection based Bayesian reconstruction of visual image from human brain activity. *Neurocomputing*, *316*, 202–209.

Huang, W., Yan, H., Wang, C., Li, J., Yang, X., Li, L., … Chen, H. (2020). Long short-term memory-based neural decoding of object categories evoked by natural images. *Human Brain Mapping*, *41*(15), 4442–4453.

Huang, W., Yan, H., Wang, C., Li, J., Zuo, Z., Zhang, J., … Chen, H. (2020). Perception-to-image: Reconstructing natural images from the brain activity of visual perception. *Annals of Biomedical Engineering*, *48*, 2323–2332. https://doi.org/10.1007/s10439-020-02502-3

Huang, W., Yan, H., Wang, C., Yang, X., Li, J., Zuo, Z., … Chen, H. (2021). Deep natural image reconstruction from human brain activity based on conditional progressively growing generative adversarial networks. *Neuroscience Bulletin*, *37*, 369–373.

Huth, A. G., Lee, T., Nishimoto, S., Bilenko, N. Y., Vu, A. T., & Gallant, J. L. (2016). Decoding the semantic content of natural movies from human brain activity. *Frontiers in Systems Neuroscience*, *10*, 81.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Jack, R. E., & Schyns, P. G. (2017). Toward a social psychophysics of face communication. *Annual Review of Psychology*, *68*, 269–297.

Jafakesh, S., Jahromy, F. Z., & Daliri, M. R. (2016). Decoding of object categories from brain signals using cross frequency coupling methods. *Biomedical Signal Processing & Control*, *27*, 60–67.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv Preprint*, 1607.01759.

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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv Preprint*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain*.

Makin, J. G., Moses, D. A., & Chang, E. F. (2020). Machine translation of cortical activity to text with an encoder-decoder framework. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 708206, *23*, 575–582.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Preprint*. https://arxiv.org/pdf/1301.3781v3.pdf.

Nishida, S., & Nishimoto, S. (2018). Decoding naturalistic experiences from human brain activity via distributed representations of words. *NeuroImage*, *180*, 232–242.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar*.

Polat, K., & Güneş, S. (2009). A new feature selection method on classification of medical datasets: Kernel F-score feature selection. *Expert Systems with Applications*, *36*(7), 10367–10373.

Qiao, K., Chen, J., Wang, L., Zhang, C., Zeng, L., Tong, L., & Yan, B. (2019). Category decoding of visual stimuli from human brain activity using a bidirectional recurrent neural network to simulate bidirectional information flows in human visual cortices. *Frontiers in Neuroscience*, *13*(692), 1–15.

Song, S., Zhan, Z., Long, Z., Zhang, J., & Yao, L. (2011). Comparative study of SVM methods combined with voxel selection for object category classification on fMRI data. *PLoS One*, *6*(2), e17191.

Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, *79*(5), 1025–1034.

Taghizadeh-Sarabi, M., Daliri, M. R., & Niksirat, K. S. (2015). Decoding objects of basic categories from electroencephalographic signals using wavelet transform and support vector machines. *Brain Topography*, *28*(1), 33–46.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Proceedings of 31st Conference on Neural Information Processing Systems, Long Beach, California*. https://arxiv.org/pdf/1706.03762.pdf.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. *Proceedings of the IEEE conference on computer vision and pattern recognition, San Francisco, CA*.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *Proceedings of the IEEE conference on computer vision and pattern recognition, San Francisco, California*.

Wang, C., Yan, H., Huang, W., Li, J., Yang, J., Li, R., … Zuo, Z. (2020). "When" and "what" did you see? A novel fMRI-based visual decoding framework. *Journal of Neural Engineering*, *17*(5), 056013.

Wang, S., Zhang, J., Wang, H., Lin, N., & Zong, C. (2020). Fine-grained neural decoding with distributed word representations. *Information Sciences*, *507*, 256–272.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Huang, W., Yan, H., Cheng, K., Wang, Y., Wang, C., Li, J., Li, C., Li, C., Zuo, Z., & Chen, H. (2021). A dual-channel language decoding from brain activity with progressive transfer training. *Human Brain Mapping*, *42*(15), 5089–5100. https://doi.org/10.1002/hbm.25603