Neural Networks 170 (2024) 349–363

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Mental image reconstruction from human brain activity: Neural decoding of mental imagery via deep neural network-based Bayesian estimation

Naoko Koide-Majima [a,b,1], Shinji Nishimoto [a,b,c], Kei Majima [d,e,1,*]

[a] Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology, Osaka 565-0871, Japan
[b] Graduate School of Frontier Biosciences, Osaka University, Osaka 565-0871, Japan
[c] Graduate School of Medicine, Osaka University, Osaka 565-0871, Japan
[d] Institute for Quantum Life Science, National Institutes for Quantum Science and Technology, Chiba 263-8555, Japan
[e] JST PRESTO, Saitama 332-0012, Japan

ARTICLE INFO

Keywords:
Mental image
Brain decoding
Semantic representation
Bayesian estimation

ABSTRACT

Visual images observed by humans can be reconstructed from their brain activity. However, the visualization (externalization) of mental imagery is challenging. Only a few studies have reported successful visualization of mental imagery, and their visualizable images have been limited to specific domains such as human faces or alphabetical letters. Therefore, visualizing mental imagery for arbitrary natural images stands as a significant milestone. In this study, we achieved this by enhancing a previous method. Specifically, we demonstrated that the visual image reconstruction method proposed in the seminal study by Shen et al. (2019) heavily relied on low-level visual information decoded from the brain and could not efficiently utilize the semantic information that would be recruited during mental imagery. To address this limitation, we extended the previous method to a Bayesian estimation framework and introduced the assistance of semantic information into it. Our proposed framework successfully reconstructed both seen images (i.e., those observed by the human eye) and imagined images from brain activity. Quantitative evaluation showed that our framework could identify seen and imagined images highly accurately compared to the chance accuracy (seen: 90.7%, imagery: 75.6%, chance accuracy: 50.0%). In contrast, the previous method could only identify seen images (seen: 64.3%, imagery: 50.4%). These results suggest that our framework would provide a unique tool for directly investigating the subjective contents of the brain such as illusions, hallucinations, and dreams.

Introduction

Neural decoding technologies enable the visualization of perceptual contents based on brain activity (Kay & Gallant, 2009; Rakhimberdina et al., 2021). Previous studies have demonstrated that images seen by human participants can be reconstructed from the brain activity measured using functional magnetic resonance imaging (fMRI). Several studies have reconstructed visual perception for specific domains such as human faces (Cowen et al., 2014; H. Lee & Kuhl, 2016), hand-written letters (Schoenmakers et al., 2013), and binary images (Fujiwara et al., 2013; Miyawaki et al., 2008; Satake et al., 2018). Other studies have decoded seen natural images (Kay et al., 2008; Naselaris et al., 2009) or videos (Nishimoto et al., 2011) using visual features inspired by neurophysiological discoveries. Recently, by incorporating the assistance of deep neural networks (DNNs)(Krizhevsky et al., 2012; Simonyan & Zisserman, 2014) and generative models (Brock et al., 2019; Dhariwal & Nichol, 2021; Esser et al., 2021; Oord et al., 2018; Radford et al., 2016; Razavi et al., 2019; Rombach et al., 2022; Song et al., 2021), several studies have achieved higher-fidelity natural image reconstruction (Beliy et al., 2019; Chen et al., 2023; Fang et al., 2020; Gaziv et al., 2022; Han et al., 2019; Lu et al., 2023; Mozafari et al., 2020; Ozcelik & VanRullen, 2023; Qiao et al., 2018; Ren et al., 2021; Seeliger et al., 2018; Shen et al., 2019; St-Yves & Naselaris, 2018; Takagi & Nishimoto, 2023; VanRullen & Reddy, 2019), which has become a tool for investigating the visual processing in the brain (e.g., visual representation (Chang et al., 2023; Nestor et al., 2020), attention (Horikawa & Kamitani, 2022), and illusion (Cheng et al., 2023)).

Previous studies have succeeded in reconstructing images seen by

* Corresponding author at: Institute for Quantum Life Science, National Institutes for Quantum Science and Technology, Chiba, 263-8555, Japan.
E-mail address: majima.kei@qst.go.jp (K. Majima).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.neunet.2023.11.024
Received 10 June 2023; Received in revised form 22 September 2023; Accepted 8 November 2023
Available online 9 November 2023
0893-6080/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

humans from their brain activity; however, externalizing mental imagery remains a challenge. Only a few studies have reported successful visualization of mental imagery, and their visualizable images have been limited to specific domains such as human faces (H. Lee & Kuhl, 2016), alphabetical letters (Senden et al., 2019), or geometric shapes (Shen et al., 2019). Therefore, visualizing mental imagery for arbitrary natural images stands as a significant milestone. Among those previous studies, Shen et al. (2019) attempted to reconstruct both seen and imagined arbitrary natural images; however, they reported that only rough silhouettes of geometric shapes could be barely reconstructed from the brain activity during imagery. As we will demonstrate, one possible reason for this limitation is that this previous method heavily relied on low-level visual information decoded from the brain. According to other neuroimaging studies, high-level or semantic information (representation) is thought to be recruited more strongly in the brain during mental imagery than low-level visual information. Although low-level visual features of imagined images (e.g., Gabor-wavelet features) can be decoded to a certain extent (Albers et al., 2013; Harrison & Tong, 2009; Naselaris et al., 2015; Xing et al., 2013), high-level visual features are more helpful in identifying imagined objects from brain activity (Horikawa & Kamitani, 2017). Furthermore, categories of imagined objects can be better predicted from the brain activity in high-level visual areas than in low-level visual areas (Dijkstra et al., 2019; S.-H. Lee et al., 2012; Reddy et al., 2010). Thus, high-level and semantic information should be efficiently incorporated into the image reconstruction method to successfully externalize mental imagery.

To overcome the limitations of the method by Shen et al. (2019), we first extended this previous method to a Bayesian estimation framework and then introduced the assistance of semantic information. In the previous method, brain activity measured by fMRI was first translated (decoded) into VGG19's hierarchical representations (i.e., unit activations of individual layers in VGG19) (Simonyan & Zisserman, 2014) using a variant of linear regression (Fig. 1a). Subsequently, an image was generated using an iterative process, such that the generated image would lead to unit activations similar to those decoded from the brain. The resulting image was considered a reconstruction. Whereas all convolutional and fully-connected layers of VGG19 were combined in the previous study, as we will demonstrate, this previous method failed to produce meaningful images using only high VGG layers. Accordingly, this method relied heavily on low-level visual information. In our current study, by viewing the image generation process in this method as maximum likelihood estimation, we extended it to Bayesian estimation (Fig. 1b). This framework enables us to use a sophisticated prior of natural images developed in recent computer vision studies, which is expected to help produce meaningful images even from abstract or partial information.

In Bayesian estimation, sampling from a posterior distribution is often intractable; thus, its application in neural decoding has been limited. Although a few previous studies have introduced Bayesian estimation into letter (Schoenmakers et al., 2013) and face (Güçlütürk et al., 2017) image reconstruction, these were cases where the posterior distribution could be analytically obtained. Their approach cannot be straightforwardly applied to cases with natural images. As an alternative approach, we used the stochastic gradient Langevin dynamics (SGLD) algorithm (Welling & Teh, 2011) to sample images from the posterior distribution. Our results demonstrated that seen and imagined images can be successfully reconstructed from brain activity, supporting the effectiveness of the SGLD algorithm in the field of neural decoding.

Subsequently, a pre-trained contrastive language–image pre-training (CLIP) model (Radford et al., 2021) was used to leverage semantic information from the brain for image reconstruction. Because the CLIP model has been trained to obtain embeddings shared between images and their text captions in the last layer, the image encoder of the CLIP model is thought to extract semantic information from input images. As terminology, features and representations provided by the last layer of the CLIP model are called semantic features and semantic representations, respectively. Similarly, those provided by low/high layers of VGG19 are called low/high-level visual features and low/high-level visual representations. Here, using the same neural decoding procedure as that for VGG19, brain signals were translated into the semantic features provided by the CLIP model (Fig. 1a). Thereafter, the decoded semantic features were introduced into our Bayesian image reconstruction framework through an additional likelihood function (Fig. 1b).

By applying the proposed framework to the dataset from Shen et al. (2019), we demonstrate that our framework can reconstruct seen images only using high-level visual information and externalize mental imagery.

Related work

Neural decoding of mental imagery

Several previous studies have used machine learning classifiers to decode the contents of mental imagery. In a pioneering work from 2009, Harrison and Tong (2009) demonstrated that a classifier trained to predict the orientations of seen (i.e., observed) gratings from fMRI signals could also predict the orientations of gratings remembered in the mind during a working memory task. While this initial study focused on the contents of working memory rather than mental imagery, the same classification approach has subsequently been used to decode the categories of imagined objects and scenes (Albers et al., 2013; Cichy et al., 2012; Dijkstra et al., 2019; S.-H. Lee et al., 2012; Reddy et al., 2010; Stokes et al., 2009). As an extension, Horikawa and Kamitani (2017) constructed decoding models that predict hierarchical visual features, enabling them to identify the categories of imagined objects by applying the trained decoders to fMRI signals during imagery. A similar paradigm was recently adopted in a study involving human electrocorticographic signals (Fukuma et al., 2022). Taking a similar yet distinct approach, Naselaris et al. (2015) successfully identified imagined artworks from a given database using voxel-wise encoding models with low-level visual features.

While those previous studies have decoded the contents of mental imagery using classification or retrieval approaches, only a limited number of studies have managed to achieve the successful visualization (i.e., reconstruction) of mental imagery. Furthermore, it is important to note that the visualizable images in these studies have been confined to specific domains such as human faces (H. Lee & Kuhl, 2016), alphabetical letters (Senden et al., 2019), and geometric shapes (Shen et al., 2019). Consequently, visualizing mental imagery for arbitrary natural images remains a challenging undertaking.

Visual image reconstruction based on Bayesian estimation

A few studies have incorporated Bayesian estimation into the process of visual image reconstruction and demonstrated notable improvements in reconstruction quality. Since Bayesian estimation often involves sampling from an intractable posterior distribution, its application has been largely limited. Schoenmakers et al. (2013) solved this issue by adopting a Gaussian distribution as a prior, resulting in a posterior distribution that can be solved analytically. The same approach was also used in another study from the same group (Güçlütürk et al., 2017). As another approach, Qiao et al. (2020) circumvented this problem by selecting the image with the highest posterior probability from a synthetic image database. As an alternative approach, our work introduced the SGLD algorithm for Bayesian sampling (Welling & Teh, 2011). It should be also noted that the aforementioned previous studies focused on the reconstruction of observed images, whereas our study extends its scope to encompass both observed and imagined images.

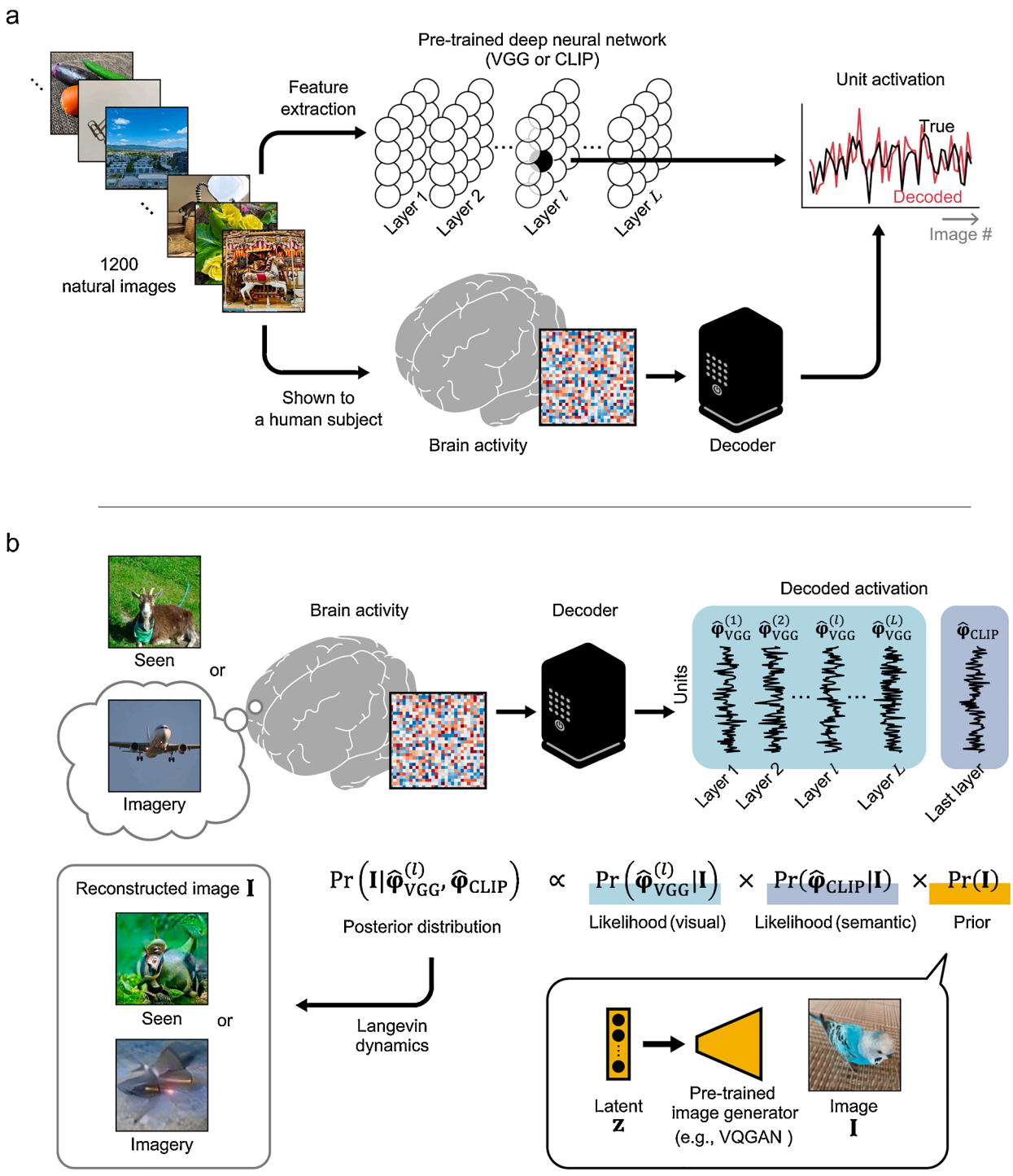

Fig. 1. Proposed reconstruction framework. (a) Decoder training. In our framework, brain activity is translated (decoded) into internal representations of a pre-trained deep neural network (DNN). Functional magnetic resonance imaging (fMRI) responses measured while a human subject viewed 1200 natural images were used as training data. Linear regression models were trained to predict unit activations of individual DNN units in the DNN that responds to the same images. The pre-trained VGG and CLIP models were used as the DNNs in this study. **(b)** Image reconstruction through Bayesian estimation. The seen or imagined image is reconstructed from the decoded DNN unit activations. The decoded DNN unit activations are incorporated into the Bayesian posterior distribution of the image via likelihood functions. The three terms on the right side of the equation correspond to Eqs. (1)-(3) in the main text, respectively. A prior distribution constructed with a pre-trained neural network-based image generator model is used in the Bayesian estimation. An image sampled from the posterior distribution is considered a reconstruction.

Methods

fMRI datasets

We used the fMRI dataset from a previous study (Shen et al., 2019), which can be downloaded from the Figshare repository (https://figshare.com/articles/dataset/Deep_Image_Reconstruction/7033577). The dataset comprised fMRI data from three human subjects: Subject 1 (male, age 33), Subject 2 (male, age 23), and Subject 3 (female, age 23). This sample size had been adopted based on previous fMRI studies with similar experimental designs (Horikawa & Kamitani, 2017; Miyawaki et al., 2008). In this experiment, each subject viewed or imagined an image in each trial, and the brain activity was measured using fMRI. The fMRI data were divided into two sets: training and test datasets. The training dataset was used for decoder training and the test dataset was used for evaluation in the previous study. The same data splitting approach was adopted in the present study.

The training dataset comprised fMRI data measured while the subjects viewed 1200 natural images. The images were collected from the online image database ImageNet (2011, fall release) (Deng et al., 2009), which were identical to those used in Horikawa and Kamitani (2017). Each image was presented to each subject five times. Thus, 6000 fMRI responses per subject were available as training data.

The test dataset comprised fMRI data measured while the subjects viewed 50 natural images and 40 artificial shapes (geometric shapes) and those measured while the subjects imagined 10 natural images and 15 artificial shapes. The 50 natural images used for this experiment were also collected from ImageNet, and they had no category overlap with the 1200 natural images used for the training dataset. The 40 artificial shapes consisted of the 40 combinations of five shapes and eight colors (red, green, blue, cyan, magenta, yellow, white, and black). The five shapes were identical to those used in Miyawaki et al. (2008). Out of those 50 natural images and 40 artificial shapes, 10 natural images and 15 artificial shapes were used as target images to be imagined. Prior to the fMRI experiment where the subjects imagined images, the subjects viewed those 25 images and remembered 25 word cues associated with the individual images. In the subsequent fMRI experiment, one of the word cues was presented in each single trial and the subjects imagined the corresponding image. For the test dataset, each subject viewed each natural image 24 times and each artificial shape 20 times, and each subject imagined each natural image 20 times and each artificial shape 20 times.

To adopt the same fMRI preprocessing procedure as in the previous study, we downloaded the preprocessed fMRI data from the Figshare repository. Following the same procedure, the training data were used without trial averaging and the trial-averaged test data were used for evaluation.

Pre-trained neural networks

Three pre-trained neural networks were used in this study: VGG19 (Simonyan & Zisserman, 2014), VQGAN (Esser et al., 2021), and CLIP's image encoder (Radford et al., 2021). We used a pre-trained VGG19 model provided by PyTorch. The outputs (unit activations) from conv1_2, conv2_2, conv3_4, conv4_4, conv5_4, fc6, fc7, and fc8 layers were used as the hierarchical representations in the brain decoding analysis. Following the procedure by Shen et al. (2019), the unit activation values before rectification were used as the targets to be decoded. In this study, these eight layers are called conv1, conv2, conv3, conv4, conv5, fc6, fc7, and fc8, and the unit activation vector of the l-th layer for an input image $\mathbf{I} \in \mathbb{R}^{224\times224\times3}$ is denoted by $\boldsymbol{\varphi}^{(l)}_{\mathrm{VGG}}(\mathbf{I})$.

A pre-trained VQGAN model was downloaded from the official GitHub repository (https://github.com/CompVis/taming-transformers). The model "VQGAN ImageNet ($f = 16$), 1024″ was used in this study. VQGAN uses a latent vector $\mathbf{z}$ as the input and produces an image as the output. The probability distribution of the output image given $\mathbf{z}$ is denoted by $p_{\mathrm{VQGAN}}(\mathbf{I}|\mathbf{z})$. Note that the output of VQGAN is deterministic with respect to $\mathbf{z}$. However, we describe our framework using a probability distribution because our framework can also be combined with image generator models that probabilistically produce images.

The CLIP image encoder was downloaded from the official GitHub repository (https://github.com/openai/CLIP). The model "ViT-B/32″ was used in this study. The output from the last layer was used as the target to be decoded and denoted by $\boldsymbol{\varphi}_{\mathrm{CLIP}}(\mathbf{I})$.

Conventionally, the features and representations provided by low/high layers of VGG19 are called low/high-level visual features and low/high-level visual representations. Similarly, those provided by the last layer of CLIP are called semantic features and semantic representations in this study.

Brain decoder

In our proposed framework, brain activity measured using fMRI was translated (decoded) into hierarchical representations of VGG19 (Fig. 1a). For decoder construction, linear regression models were trained to predict the unit activations of individual units in each layer of VGG19 using the training dataset (i.e., fMRI responses to 1200 natural images). We used the linear regression algorithm with L2-regularization. Unless stated otherwise, fMRI signals from the voxels in the whole visual cortex were used as input for predicting the layers with spatial dimensions (i.e., conv1–conv5), because all individual subareas in the visual cortex are known to have considerable spatial information (Majima et al., 2017). To predict the layers without spatial dimensions (i.e., fc6–fc8), fMRI signals from the voxels in the higher visual cortex (HVC) were used. According to previous studies, these layers can be accurately predicted from fMRI signals in HVC, and this choice is expected to reduce the risk of overfitting (Horikawa & Kamitani, 2017; Nonaka et al., 2021). Before linear regression training, fMRI voxels (i.e., input dimensions) were selected using the following voxel selection procedure.

Input voxel selection was performed using the training dataset to reduce the computational time and the risk of overfitting. To predict a given VGG layer, we applied principal component analysis (PCA) to its representations across the 1200 images and extracted the principal components that explained more than 99% of the variance. Subsequently, for each voxel, we computed the correlation coefficients between the fMRI signal and the individual principal components. The maximum absolute value of the correlation coefficients was assigned to the voxel. This procedure was repeated for all voxels, and the voxels were ranked in descending order of the assigned correlation values. The top k voxels were used as inputs for the L2-regularized linear regression algorithm. The activation of each unit in the layer was predicted from the fMRI signals in the selected voxels. The algorithm of this process was summarized as pseudocode Algorithm 1. The number of voxels used (k) and the regularization parameter (λ) were optimized using the training dataset. In the optimization process, 20% of the training data were randomly chosen as validation data, a linear regression model was trained using the rest of the training data, and the trained model was tested on the validation data. k and λ were varied across $\{1000, 2000, \cdots, 5000\}$ and $\{2^2, 2^4, 2^6, \cdots, 2^{18}\}$, respectively. The hyperparameter values leading to the highest prediction performance on the validation data were adopted. The algorithm of this hyperparameter optimization process was summarized as pseudocode Algorithm 2.

The trained linear regression models (brain decoders) were then applied to fMRI data in the test dataset. Following the procedure of Shen et al. (2019), this step and the subsequent image reconstruction process were applied to the trial-averaged fMRI response (i.e., trial-averaged fMRI activity pattern) for each image in each condition. For a given trial-averaged activity pattern, the decoded VGG representations are denoted by $\widehat{\varphi}^{(l)}_{\mathrm{VGG}}$ $(l = 1, \cdots 8)$ in this study. The same decoding procedure

Algorithm 1
Decoder training for each DNN layer.

Input: fMRI data $\mathbf{X} \in \mathbb{R}^{N\times D}$, unit activations in a DNN layer $\mathbf{\Phi} \in \mathbb{R}^{N\times D'}$, the number of input voxels for L2-regularized linear regression k, the value of the regularization parameter λ
Output: Linear weight $\mathbf{W} \in \mathbb{R}^{D\times D'}$
1: Apply PCA to $\mathbf{\Phi}$ and extract PCs that explain more than 99% of the variance.
Denote the resultant PCs by $\mathbf{Z} \in \mathbb{R}^{N\times D'}$.
2: Compute the correlation coefficients $\mathbf{C} = \mathrm{corr}(\mathbf{X}, \mathbf{Z}) \in \mathbb{R}^{D\times D'}$.
3: Compute the absolute values of the individual elements of $\mathbf{C}$ as $\mathrm{abs}(\mathbf{C}) \in \mathbb{R}^{D\times D'}$.
4: Compute $\max(\mathrm{abs}(\mathbf{C}),\ \mathrm{axis} = 2) \in \mathbb{R}^{D}$.
5: Sort the elements of $\max(\mathrm{abs}(\mathbf{C}),\ \mathrm{axis} = 2)$.
Denote the indices of the k dimensions with the highest values by $S \subset \{1, \cdots, D\}$.
6: For $d' = 1$ to D' do
7: Apply L2-reguralized linear regression to $\mathbf{X}(:, S)$ and $\mathbf{\Phi}(:, d')$ with the regularization parameter set to λ.
Back-project the obtained weight into the original D-dimensional space.
Denote the results by $\mathbf{W}(:, d') \in \mathbb{R}^{D}$.
8: End for
9: Concatenate $\mathbf{W}(:, d')$ $(d' = 1, \cdots, D')$ and denote the results by $\mathbf{W} \in \mathbb{R}^{D\times D'}$.

Algorithm 2
Hyperparameter optimization.

Input: fMRI data $\mathbf{X} \in \mathbb{R}^{N\times D}$, unit activations in a DNN layer $\mathbf{\Phi} \in \mathbb{R}^{N\times D'}$
Output: Optimized hyperparameters $\hat{k}$ and $\hat{\lambda}$
1: $k_1, k_2, \cdots, k_5 \leftarrow 1000,\ 2000, \cdots,\ 5000$
2: $\lambda_1, \lambda_2, \cdots, \lambda_9 \leftarrow 2^2,\ 2^4, \cdots,\ 2^{18}$
3: Randomly split the rows of $\mathbf{X}$ into 20% and 80%.
Denote the corresponding matrices by $\mathbf{X}_{\mathrm{val}}$ and $\mathbf{X}_{\mathrm{train}}$.
4: Split the rows of $\mathbf{\Phi}$ with the same splitting.
Denote the corresponding matrices by $\mathbf{\Phi}_{\mathrm{val}}$ and $\mathbf{\Phi}_{\mathrm{train}}$.
5: for $i = 1$ to 5 do
6: for $j = 1$ to 9 do
7: $\mathbf{W} \leftarrow \mathrm{Algorithm1}(\mathbf{X}_{\mathrm{train}}, \mathbf{\Phi}_{\mathrm{train}}, k_i, \lambda_j)$ (Decoder training for each DNN layer)
8: $\hat{\mathbf{\Phi}} \leftarrow \mathbf{X}_{\mathrm{val}}\ \mathbf{W}$
9: Compute the correlation coefficients $\mathbf{C} = \mathrm{corr}(\mathbf{\Phi}_{\mathrm{val}}, \hat{\mathbf{\Phi}}) \in \mathbb{R}^{D'\times D'}$.
10: Compute mean(diag($\mathbf{C}$)). Denote the results by $r_{i,j}$.
11: End for
12: End for
13: $\hat{i}, \hat{j} \leftarrow \underset{i,j}{\mathrm{argmax}}\{r_{i,j}\}$
14: $\hat{k}, \hat{\lambda} \leftarrow k_{\hat{i}}, \lambda_{\hat{j}}$

was also performed to predict the last layer of CLIP, and the decoded representation is denoted by $\hat{\varphi}_{\mathrm{CLIP}}$. The algorithm of the full decoder training for all used layers was summarized as pseudocode Algorithm 3, and the procedure of applying the trained decoders to a given fMRI activity pattern was summarized as pseudocode Algorithm 4.

In the analysis shown in Fig. 5, to compare the reconstruction quality between subareas in the visual cortex, we performed the above decoding procedure using only fMRI signals in each of V1, V2, V3, V4, and HVC. We selected the voxels in each subarea using the labels provided with the preprocessed fMRI data from the Figshare repository.

Algorithm 3
Decoder training (Full).

Input: fMRI data $\mathbf{X} \in \mathbb{R}^{N\times D}$, unit activations of VGG $\mathbf{\Phi}^{(l)}_{\mathrm{VGG}}$ $(l = 1, \cdots, L)$, unit activations of CLIP $\mathbf{\Phi}_{\mathrm{CLIP}}$
Output: Linear weights $\mathbf{W}^{(l)}_{\mathrm{VGG}}$ $(l = 1, \cdots, L)$, $\mathbf{W}_{\mathrm{CLIP}}$
1: For $l = 1$ to L do
2: $[\hat{k}, \hat{\lambda}] \leftarrow \mathrm{Algorithm2}(\mathbf{X}, \mathbf{\Phi}^{(l)}_{\mathrm{VGG}})$ (Hyperparameter optimization)
3: $\mathbf{W}^{(l)}_{\mathrm{VGG}} \leftarrow \mathrm{Algorithm1}(\mathbf{X}, \mathbf{\Phi}^{(l)}_{\mathrm{VGG}}, \hat{k}, \hat{\lambda})$ (Decoder training for each DNN layer)
4: End for
5: $[\hat{k}, \hat{\lambda}] \leftarrow \mathrm{Algorithm2}(\mathbf{X}, \mathbf{\Phi}_{\mathrm{CLIP}})$ (Hyperparameter optimization)
6: $\mathbf{W}_{\mathrm{CLIP}} \leftarrow \mathrm{Algorithm1}(\mathbf{X}, \mathbf{\Phi}_{\mathrm{CLIP}}, \hat{k}, \hat{\lambda})$ (Decoder training for each DNN layer)

Algorithm 4
Brain decoding.

Input: fMRI data $\mathbf{x} \in \mathbb{R}^{D}$, Linear weights $\mathbf{W}^{(l)}_{\mathrm{VGG}}$ $(l = 1, \cdots, L)$, $\mathbf{W}_{\mathrm{CLIP}}$
Output: Decoded unit activations $\hat{\varphi}^{(l)}_{\mathrm{VGG}}$, $\hat{\varphi}_{\mathrm{CLIP}}$
1: For $l = 1$ to L do
2: $\hat{\varphi}^{(l)}_{\mathrm{VGG}} \leftarrow \mathbf{W}^{(l)\ \mathrm{T}}_{\mathrm{VGG}}\ \mathbf{x}$
3: End for
4: $\hat{\varphi}_{\mathrm{CLIP}} \leftarrow \mathbf{W}^{\mathrm{T}}_{\mathrm{CLIP}}\ \mathbf{x}$

Proposed reconstruction framework

This section describes the proposed reconstruction algorithm. The seen or imagined image was reconstructed from $\hat{\varphi}^{(l)}_{\mathrm{VGG}}$ and $\hat{\varphi}_{\mathrm{CLIP}}$ using a Bayesian estimation framework.

We assume that the log likelihood functions are given by

$$\log p\left(\hat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \hat{\varphi}^{(L)}_{\mathrm{VGG}} \middle| \mathbf{I}\right) = \frac{1}{T}\sum_{l=1}^{L} w_l\ \mathrm{similarity}\left(\hat{\varphi}^{(l)}_{\mathrm{VGG}}, \boldsymbol{\varphi}^{(l)}_{\mathrm{VGG}}(\mathbf{I})\right) +\ \mathrm{const}, \tag{1}$$

and

$$\log p(\hat{\varphi}_{\mathrm{CLIP}}|\mathbf{I}) = \frac{\lambda_{\mathrm{CLIP}}}{T}\int \mathrm{similarity}(\hat{\varphi}_{\mathrm{CLIP}}, \boldsymbol{\varphi}_{\mathrm{CLIP}}(\mathbf{I}'))\ p_{\mathrm{Perturb}}(\mathbf{I}'|\mathbf{I})d\mathbf{I}' +\ \mathrm{const}\ . \tag{2}$$

Here, T is a parameter called "temperature", and was set to 10^{-7}. The function $\mathrm{similarity}(\cdot, \cdot)$ is a similarity metric to measure the similarity between two input vectors. The Pearson correlation coefficient was used as the similarity metric in this study. When the negative L2 norm is adopted as the similarity metric, the maximum likelihood estimation using this likelihood function is equivalent to the reconstruction method proposed by Shen et al. (2019). Thus, our proposed image reconstruction method is a Bayesian extension of that of Shen et al. (2019). In Eq. (1), w_l is a parameter that controls the strength of the contribution from the l-th VGG layer, and this was set to the inverse of the number of used VGG layers. The reconstruction was performed using each VGG layer, using a subset of the VGG layers, or using all VGG layers. The sum in Eq. (1) was taken across the used layer(s). For Eq. (2), λ_{CLIP} is a parameter that controls the strength of the semantic assistance. It was set to 0.1 and 0.25 for seen image reconstruction and imagery reconstruction, respectively. To make the likelihood function robust to perturbations in $\mathbf{I}$, a random affine transformation is applied to $\mathbf{I}$, and the mean similarity is evaluated in the likelihood function. The conditional probability distribution $p_{\mathrm{Perturb}}(\mathbf{I}'|\mathbf{I})$ represents the random affine transformation, and we set its details according to the settings adopted in an implementation of text-to-image generation (https://medium.com/geekculture/text-to-image-synthesis-using-multimodal-vqgan-clip-architectures-896b8a6588ef). When we use the SGLD algorithm, the mean in Eq. (2) is replaced by the empirical mean with 32 random samples.

We used a pre-trained image generator model to prepare a prior distribution for the Bayesian estimation. Specifically, VQGAN (Esser et al., 2021) trained on ImageNet training images (Deng et al., 2009) was used in this study; however, other image generator models can be used in our framework. Given its latent vector $\mathbf{z} \in \mathbb{R}^{14\times 14\times 256}$, the image generator model of VQGAN produces an image. We denote the probability distribution of the generated images conditioned on $\mathbf{z}$ by $p_{\mathrm{VQGAN}}(\mathbf{I}|\mathbf{z})$. Note that the produced image $\mathbf{I}$ is deterministic with respect to $\mathbf{z}$ when we use VQGAN; however, here, we explain our framework with a probability distribution because our framework can also be combined with image generator models that probabilistically generate images.

We can obtain an image prior $p(\mathbf{I})$ by preparing a distribution $p(\mathbf{z})$, constructing the joint distribution $p(\mathbf{I}, \mathbf{z}) = p_{\mathrm{VQGAN}}(\mathbf{I}|\mathbf{z})p(\mathbf{z})$, and

marginalizing out $\mathbf{z}$:

$$p(\mathbf{I}) = \int p_{\mathrm{VQGAN}}(\mathbf{I}|\mathbf{z})p(\mathbf{z})d\mathbf{z}\ . \tag{3}$$

We used the non-informative distribution as $p(\mathbf{z})$ in this study. Any analytically differentiable distribution $p(\mathbf{z})$ can be used in this framework.

Taken together, our posterior distribution was given by

$$\begin{aligned} &p\left(\mathbf{I}\middle|\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}}, \widehat{\varphi}_{\mathrm{CLIP}}\right) \\ &\propto p\left(\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}}\middle|\mathbf{I}\right) \times p(\widehat{\varphi}_{\mathrm{CLIP}}|\mathbf{I}) \times p(\mathbf{I})\ . \end{aligned} \tag{4}$$

In our framework, an image is sampled from this posterior distribution and the obtained image is treated as a reconstructed image.

To obtain an image from the posterior distribution, we used the SGLD algorithm (Welling & Teh, 2011). The proposed posterior distribution can be rewritten as follows:

$$\begin{aligned} &p\left(\mathbf{I}\middle|\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}},\ \widehat{\varphi}_{\mathrm{CLIP}}\right) \\ &\quad = \int p_{\mathrm{VQGAN}}(\mathbf{I}|\mathbf{z})\ p\left(\mathbf{z}\middle|\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}},\ \widehat{\varphi}_{\mathrm{CLIP}}\right)\ d\mathbf{z}. \end{aligned} \tag{5}$$

Thus, if we can sample $\mathbf{z}$ from $p(\mathbf{z}\left|\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}},\ \widehat{\varphi}_{\mathrm{CLIP}}\right)$, we can sample an image from the above distribution through $p_{\mathrm{VQGAN}}(\mathbf{I}|\mathbf{z})$. We performed sampling from $p(\mathbf{z}\left|\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}},\ \widehat{\varphi}_{\mathrm{CLIP}}\right)$ using the SGLD algorithm. The update rule of the algorithm is given by

$$\mathbf{z}_{t+1} \quad = \mathbf{z}_t + \frac{\varepsilon_t}{2}\nabla \log p\left(\mathbf{z}_t\middle|\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}}, \widehat{\varphi}_{\mathrm{CLIP}}\right) + \boldsymbol{\eta}_t, \tag{6}$$

where $\mathbf{z}_t$ and ε_t are the sample and the learning rate at step t, and $\boldsymbol{\eta}_t$ is a vector whose elements are sampled from the Gaussian distribution $N(0, \varepsilon_t)$. In this study, the learning rate was gradually decreased, and its sequence was given by

$$\varepsilon_t = \frac{a}{(b+t)^{\gamma}}, \tag{7}$$

where a, b, and γ were set to 0.00015, 0.15, and 0.055. We performed 500 iterations, and a reconstructed image was produced using $\mathbf{z}_{500}$. For fast convergence, first, we performed the Adam algorithm with an initial random latent vector. Then, the SGLD algorithm was started using the result from the Adam algorithm as the initial latent vector.

At every update in the SGLD algorithm, we approximately computed the right-side of Eq. (6). The logarithm of $p(\mathbf{z}\left|\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}},\ \widehat{\varphi}_{\mathrm{CLIP}}\right)$ was represented by

$$\log p\left(\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}}\middle|\mathbf{z}\right) + \log p(\widehat{\varphi}_{\mathrm{CLIP}}|\mathbf{z}) + \log p(\mathbf{z}) + \mathrm{const}\ . \tag{8}$$

To approximately compute the first term, we rewrote the first term as

$$\begin{aligned} &\log p\left(\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}}\middle|\mathbf{z}\right) \\ &\quad = \log \int p\left(\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}}\middle|\mathbf{I}\right)\ p_{\mathrm{VQGAN}}(\mathbf{I}|\mathbf{z})\ d\mathbf{I} \\ &\quad = \log \mathbb{E}_{p_{\mathrm{VQGAN}}(\mathbf{I}|\mathbf{z})}\left[p\left(\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}}\middle|\mathbf{I}\right)\right]\ . \end{aligned} \tag{9}$$

Using Jensen's inequality, this term was approximated by

$$\mathbb{E}_{p_{\mathrm{VQGAN}}(\mathbf{I}|\mathbf{z})}\left[\log p\left(\widehat{\varphi}^{(1)}_{\mathrm{VGG}}, \cdots, \widehat{\varphi}^{(L)}_{\mathrm{VGG}}\middle|\mathbf{I}\right)\right]\ . \tag{10}$$

The inside of the square brackets is equivalent to Eq. (1); therefore, Eq. (10) was approximately computed by replacing the expectation with the empirical mean. The gradient of Eq. (10) could also be approximately computed using the back-propagation algorithm. The same approximation procedure could be applied to the second term in Eq. (8). With these approximated gradients, the update procedure shown in Eq. (6) was performed and the reconstructed image was obtained in our reconstruction framework. This image reconstruction process was summarized as pseudocode Algorithm 5, and the full proposed framework including decoder training, brain decoding, and this image reconstruction process was summarized as pseudocode Algorithm 6.

Reconstruction algorithm proposed by Shen et al. (2019)

The reconstructed images obtained using our framework were compared to those obtained using the method proposed by Shen et al. (2019). Using the decoded VGG representations from L layers, the observed or imagined image was reconstructed by solving the following optimization problem:

$$\widehat{\mathbf{I}} = \underset{\mathbf{I}\in\mathbb{R}^{224\times224\times3}}{\mathrm{argmin}} \sum_{l=1}^{L} w_l \| \widehat{\varphi}^{(l)}_{\mathrm{VGG}} - \boldsymbol{\varphi}^{(l)}_{\mathrm{VGG}}(\mathbf{I}) \|_2^2, \tag{11}$$

This optimization problem was solved using the momentum gradient method with some constraints in the previous study. As noted in Sub-Section 3.4, this method is equivalent to the maximum likelihood estimation using the negative L2 norm as the similarity metric in Eq. (1). We downloaded a Python implementation of their method and their decoded features from the official GitHub and Figshare repositories (https://github.com/KamitaniLab/DeepImageReconstruction, https://figshare.com/articles/dataset/Deep_Image_Reconstruction/7033577). All reconstruction results obtained using the method of Shen et al. (2019) in this study were obtained using these resources.

Evaluation of reconstruction quality

Reconstructed images were evaluated with the Inception score and pairwise image identification analysis. The Inception score is a metric often used to assess the quality of a set of synthesized images (Salimans et al., 2016). We used PyTorch's official implementation and calculated the score for each condition using the set of reconstructed natural images without data splitting (50 images in the case of seen image reconstruction and 10 images in the case of mental imagery reconstruction).

To examine whether reconstructed images preserved information about the original (i.e., target) images, we performed a pairwise image identification analysis. Following the procedures in previous studies Cowen et al., 2014; Shen et al., 2019), we examined whether each reconstructed image was more similar to the corresponding target image than another candidate image randomly selected from the test dataset. This pairwise identification was performed for all possible image pairs within four types of image sets: 1) the 50 natural images for the seen image reconstruction, 2) the 40 artificial shapes for the seen image reconstruction, 3) the 10 natural images for the mental imagery reconstruction, and 4) the 15 artificial shapes for the mental imagery reconstruction. The weighted similarity used in each reconstruction algorithm was used as the similarity metric (i.e., the sum of Eqs. (1) and (2) for the proposed framework and Eq. (11) for the method of Shen et al. (2019)). Since the exact computation of Eq. (2) is computationally infeasible, the mean in Eq. (2) was replaced by the empirical mean with 320 random samples. The proportion of correct answers was calculated for each

Algorithm 5
Image reconstruction.

Input: Decoded unit activations $\widehat{\varphi}^{(l)}_{\mathrm{VGG}}$ $(l = 1, \cdots, L)$, $\widehat{\varphi}_{\mathrm{CLIP}}$
Output: Reconstructed image $\widehat{\mathbf{I}} \in \mathbb{R}^{224\times224\times3}$
1: Initialize $\mathbf{z} \in \mathbb{R}^{14\times14\times256}$ with random values.
2: Update $\mathbf{z}$ using Adam and the objective function Eq. (8).
3: For $t = 1$ to 500 do
4: Update $\mathbf{z}$ with Eq. (6).
5: End for
6: Sample an image from $p_{\mathrm{VQGAN}}(\mathbf{I}|\mathbf{z})$. Denote the result by $\widehat{\mathbf{I}}$.

Algorithm 6
Full proposed framework.

Input: fMRI data (training) $\mathbf{X} \in \mathbb{R}^{N \times D}$, DNN unit activations (training) $\mathbf{\Phi}_{\mathrm{VGG}}^{(l)}$ $(l = 1, \cdots, L)$, $\mathbf{\Phi}_{\mathrm{CLIP}}$, fMRI data (test) $\mathbf{x} \in \mathbb{R}^{D}$
Output: Reconstructed image $\widehat{\mathbf{I}} \in \mathbb{R}^{224 \times 224 \times 3}$
1: $[\{\mathbf{W}_{\mathrm{VGG}}^{(l)}\}_{l=1}^{L}, \mathbf{W}_{\mathrm{CLIP}}] \leftarrow$ Algorithm3$(\mathbf{X}, \{\mathbf{\Phi}_{\mathrm{VGG}}^{(l)}\}_{l=1}^{L}, \mathbf{\Phi}_{\mathrm{CLIP}})$ (Decoder training (full))
2: $[\{\widehat{\varphi}_{\mathrm{VGG}}^{(l)}\}_{l=1}^{L}, \widehat{\varphi}_{\mathrm{CLIP}}] \leftarrow$ Algorithm4$(\mathbf{x}, \{\mathbf{W}_{\mathrm{VGG}}^{(l)}\}_{l=1}^{L}, \mathbf{W}_{\mathrm{CLIP}})$ (Brain decoding)
3: $\widehat{\mathbf{I}} \leftarrow$ Algorithm5$(\{\widehat{\varphi}_{\mathrm{VGG}}^{(l)}\}_{l=1}^{L}, \widehat{\varphi}_{\mathrm{CLIP}})$ (Image reconstruction)

reconstructed image, and the mean and standard deviation of the accuracy across reconstructed images were reported.

Quantification of the strength of line components in reconstructed images

The strengths of the line components for each orientation in reconstructed images were evaluated. We used the procedure described in previous studies (Cheng et al., 2023; Jafari-Khouzani & Soltanian-Zadeh, 2005). We converted a given reconstructed image into grayscale, subtracted the mean pixel value from the individual pixels, and squared the pixel values. Subsequently, the Radon transform was applied to the resulting image. The result was denoted by $R(r, \theta)$, and following the previous studies, the strength of line components with an orientation of θ was quantified by $\mathrm{Var}_r[R(r,\theta)]$. These values are shown in Fig. 4b.

Results

Reconstruction of seen images

We used experimental data from Shen et al. (2019) in this study (see Section 3.1). To assess the effectiveness of the proposed framework, we first applied it to the fMRI signals measured while the three human subjects viewed 50 natural images and 40 artificial shapes. The decoders were trained using independent training data comprising fMRI signals measured while the same subjects viewed other 1200 natural images (see Section 3.3). Images reconstructed from the brain via the set of all VGG layers, via the set of conv2–fc6, and via the individual VGG layers are shown in Fig. 2a and b; Supplementary Figs. 1–3. The iterative image generation process in the SGLD algorithm is also shown in Supplementary Mov 1. These images were compared to those reconstructed using the method described by Shen et al. (2019). Whereas both methods produced moderately nice reconstructions using the full set of VGG layers, the method of Shen et al. (2019) failed to produce meaningful images without conv1, implying its reliance on low-level visual information; in contrast, meaningful images were obtained using the proposed framework.

Note that successful reconstructions were obtained even for artificial shapes, although the brain decoders were trained using only the brain responses to natural images (Fig. 2a, right; Supplementary Figs. 4–6). These results demonstrate that our reconstruction framework has a strong generalization ability for images in a new unknown domain, excluding the possibility that it generates images by virtually picking them from limited exemplars.

We also evaluated how accurately the unit activations in individual layers were decoded from the brain. We computed the correlation coefficient between the true and decoded unit activations across the 50 natural images for each unit, and then computed the mean correlation coefficient across the units in each layer (Fig. 2c). As in previous studies (Horikawa & Kamitani, 2017; Shen et al., 2019), all individual layers were decoded with moderate accuracy. The results suggest that the decoded unit activations of high VGG layers as well as low VGG layers carry a significant amount of information on the presented images.

To quantitatively evaluate the quality of the reconstructed images, we performed two types of evaluations (see Section 3.6). First, we computed the Inception score (Salimans et al., 2016) for these images to evaluate their visual quality (Fig. 2d). Our reconstructions demonstrated higher Inception scores than those of Shen et al. (2019). Second, to examine whether the reconstructed images preserved information about the seen images, we performed a pairwise image identification analysis. Following the procedures in previous studies (Cowen et al., 2014; Shen et al., 2019), we examined whether each reconstructed image was more similar to the corresponding seen image than a randomly selected one; and reported the proportions of correct answers (Fig. 2e). The weighted similarity in each reconstruction algorithm was used as the similarity metric. Overall, the identification accuracies of the proposed framework were higher than those of Shen et al. (2019). When conv2–fc6 were used, the mean identification accuracy was 90.7±13.5% with our proposed framework and 64.3±30.6% with the method of Shen et al. (2019) (mean±SD across subjects and images). A similar tendency was consistently observed with the reconstructed images of the artificial shapes (Supplementary Fig. 7; 89.6±8.9% for our proposed framework and 66.7±32.9% for the method of Shen et al. (2019)).

Reconstruction of imagined images

We applied the reconstruction methods to the fMRI signals measured during imagery (Fig. 3a and b; Supplementary Figs. 8 and 9; also see Supplementary Mov 2 for the iterative image generation process). We used the data measured while the subjects imagined 10 natural images and 15 artificial shapes. Although the reconstruction quality varied significantly across samples, our reconstruction framework successfully produced interpretable images that reflected the target images to be imagined; in contrast with the method of Shen et al. (2019). A comparison of the decoding accuracy between the DNN layers demonstrated that the decoding accuracies for conv1 and conv2 were considerably low compared to those of the other layers (Fig. 3c), which is consistent with the view that our proposed framework leverages high-level visual information better than that proposed by Shen et al. (2019).

We evaluated the quality of the reconstructed images using the same procedure as that used for seen image reconstruction. Our proposed framework outperformed that of Shen et al. (2019) in terms of both Inception score and identification accuracy (Fig. 3d and e). The mean identification accuracy with conv2–fc6 was 75.6±22.3% for our proposed framework and 50.4±32.2% for the method of Shen et al. (2019) (mean±SD across subjects and images). Furthermore, our framework successfully reconstructed artificial shapes, although the brain decoders were trained using only brain responses to natural images (Supplementary Figs. 9 and 10; 72.9±23.2% for our proposed framework and 51.4±31.8% for the method of Shen et al. (2019)), indicating its strong generalization ability.

Interestingly, we found that, for some artificial shapes, line components in imagery reconstructions were emphasized compared to those in seen image reconstructions. A comparison of the reconstructed images for an X-shaped geometric pattern between the imagery and the seen image reconstructions is shown in Fig. 4a. Lines with orientations of approximately 45° and 135° appear in the imagery reconstructions, while rough silhouettes of the target shape were emphasized in the seen image reconstructions. This tendency was consistently observed across the three subjects and five different colors (Supplementary Fig. 11). To quantitatively assess this tendency, we quantified how strongly the line components with each orientation were included in the reconstructed images. Briefly, following the procedure described in previous studies (F. Cheng et al., 2023; Jafari-Khouzani & Soltanian-Zadeh, 2005), the strengths of the line components for individual orientations in a reconstructed image were evaluated by applying the Radon transform to the image (see Section 3.7). The imagery reconstructions had stronger line components at orientations of 45° and 135° than the seen image reconstructions (Fig. 4b). These results may reflect the sharpening effect caused by the top-down process in the brain (Abdelhack & Kamitani, 2018).

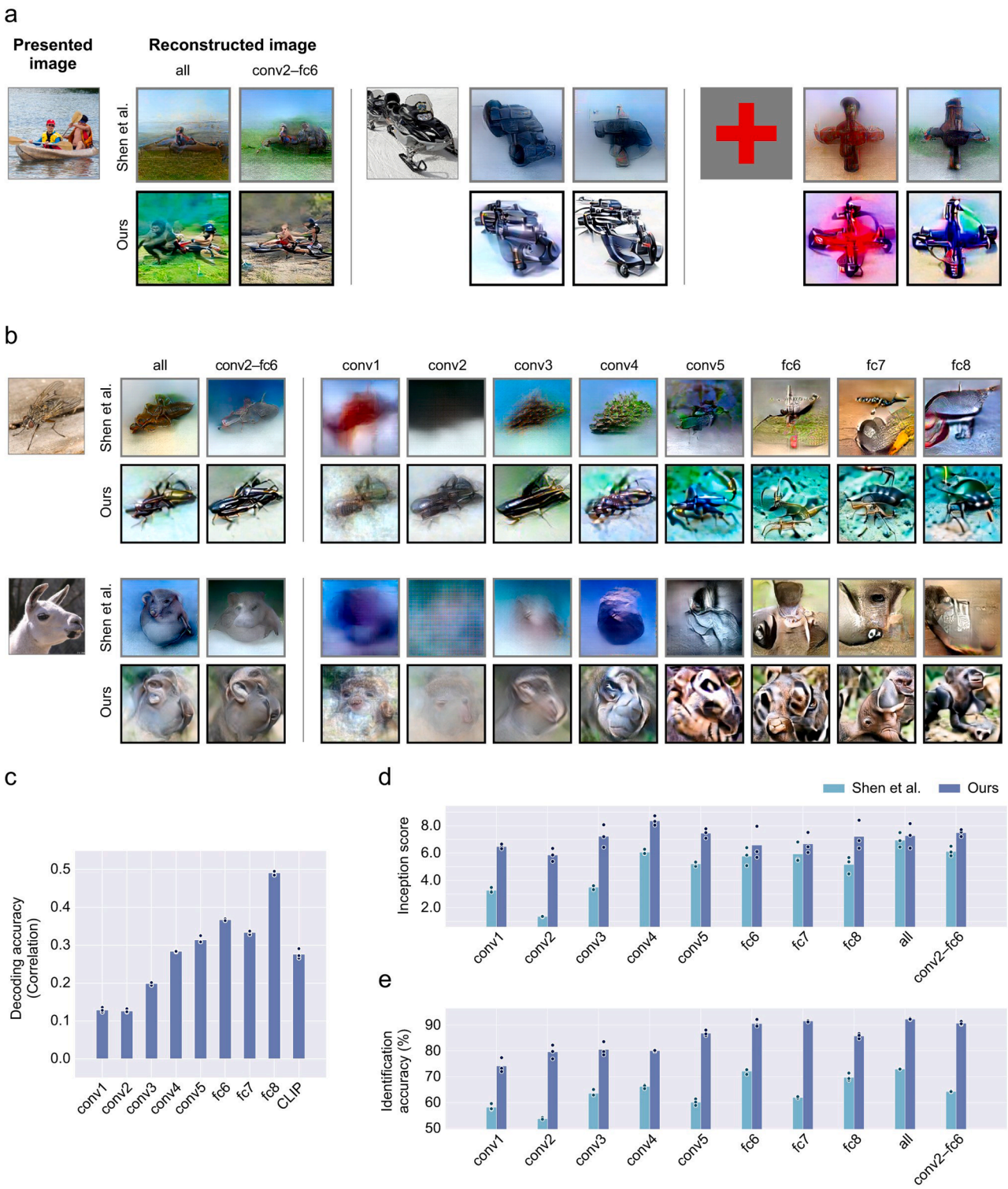

Fig. 2. Reconstruction of seen images. (a,b) Reconstructed images. Seen stimulus images were reconstructed from brain-decoded unit activations. Images reconstructed through the set of all VGG layers, through the set of conv2–fc6, and through the individual VGG layers are shown here. The gray and black surrounding frames indicate images reconstructed using the method of Shen et al. (2019) and our proposed framework, respectively. All reconstructed images shown here are from Subject 2. Reconstructed images from individual subjects are provided in Supplementary Figs. 1–3. **(c)** Decoding accuracy. The decoding accuracy for each DNN unit was evaluated by computing the correlation coefficient between true and decoded unit activation values across 50 natural images. The mean accuracy across the DNN units in each layer is shown here. Black dots and blue bars indicate the mean accuracies for individual subjects and those across the three subjects, respectively. **(d)** Inception score. Light and dark blue bars indicate the mean Inception scores across the three subjects for images reconstructed using the method of Shen et al. (2019) and our proposed framework, respectively. Dots indicate the Inception scores for individual subjects. **(e)** Image identification accuracy. The formats are the same as those in (d).

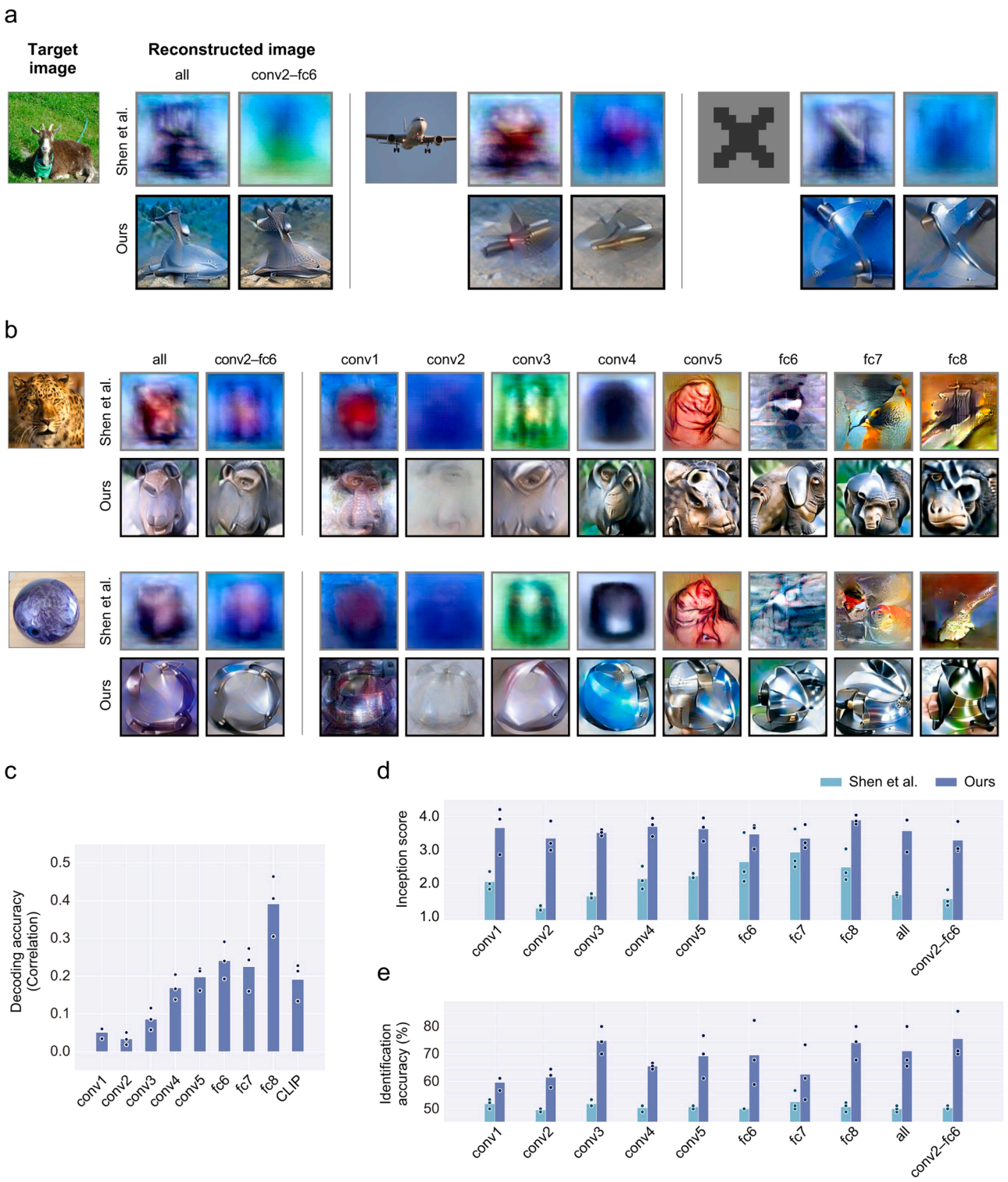

Fig. 3. Reconstruction of imagined images. Imagined images were reconstructed from brain activity. The formats are the same as those in Fig. 2. All reconstructed images in panels (a) and (b) are from Subject 2. Those from the other subjects are provided in Supplementary Fig 8.

We also conducted an analysis to investigate the contribution of each brain subarea in the visual cortex to the reconstruction. Following previous studies (Nonaka et al., 2021; Shen et al., 2019), we divided the visual cortex into five subareas: V1, V2, V3, V4, and HVC. We then performed the same decoding and reconstruction procedures while limiting the input brain area to each of those five subareas (Fig. 5). HVC outperformed the other subareas, indicating that HVC made the largest contribution to imagery reconstruction.

Effect of the image prior

To characterize the effect of the image prior, we performed an ablation analysis. To reconstruct images without the image prior, we conducted maximum likelihood estimation using the likelihood function in Eq. (4) (Fig. 6a and b). The fMRI data measured while the subjects imagined natural images were used for this analysis. The appearances of the reconstructed images using our full method were significantly better

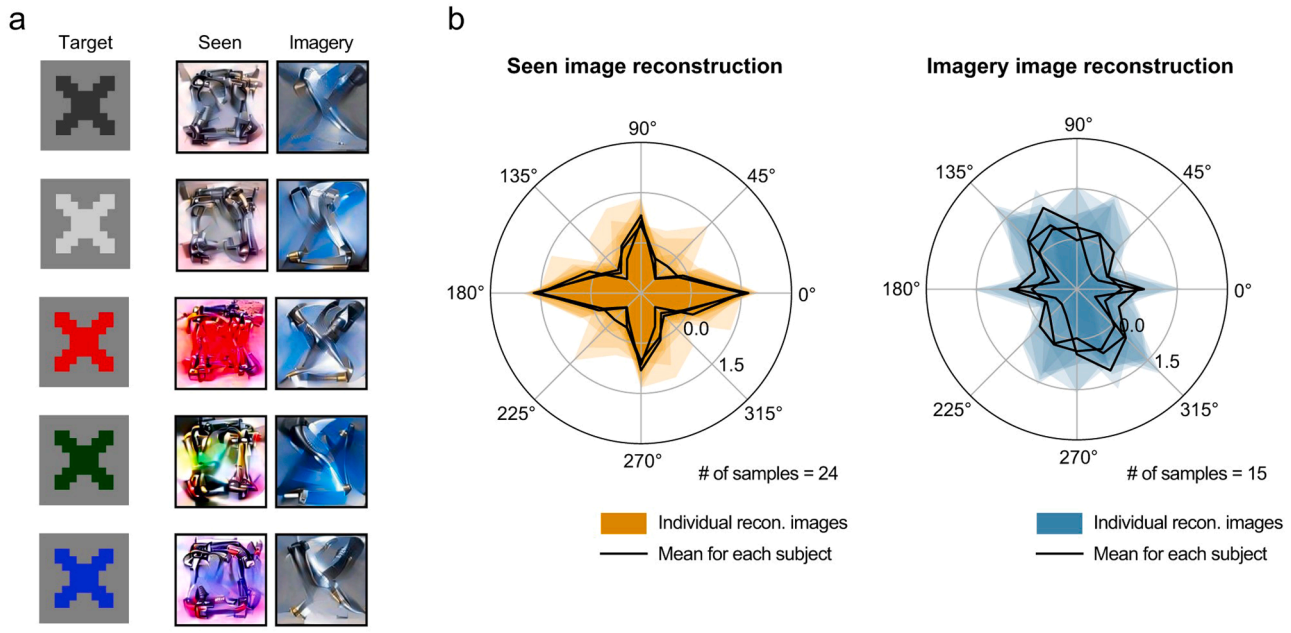

Fig. 4. Comparison of seen image reconstructions and imagery reconstructions. (a) Reconstructions of an X-shaped geometric pattern. Images reconstructed from brain activity measured while Subject 2 viewed X-shapes and images reconstructed from brain activity measured while the same subject imagined the same X-shapes are compared. All reconstructed images shown here are from Subject 2. Those from the other subjects are provided in Supplementary Fig. 11. **(b)** Quantitative evaluation. The strength of line components with each orientation in individual reconstructed images for X-shapes was evaluated as a quantitative assessment. The results for seen image reconstructions (left) and those for imagery reconstructions (right) are shown in polar plots.

than those obtained without the image prior. The quantitative comparison of the Inception score and image identification accuracy also supported this tendency (Fig. 6c and d). We also observed the same tendency with the seen image reconstructions (Supplementary Fig. 12)

Effect of the assistance of semantic information

We performed an ablation analysis to characterize the effects of the semantic information assistance. Here, we gradually varied the value of the hyperparameter controlling the strength of the influence of the CLIP features (coefficient λ_{CLIP}; see Section 3.4 for details) and reconstructed the images. $\lambda_{\mathrm{CLIP}} = 0$ indicates no assistance whereas a higher coefficient introduces stronger assistance into the reconstructed images. $\lambda_{\mathrm{CLIP}} = 0.25$ was used as the default. The imagery reconstructions were compared across different coefficient values (Fig. 7a). Removing the assistance of semantic information resulted in a drop in the image identification accuracy while the Inception score was almost maintained (Fig. 7b and c). Interestingly, small or no improvements were observed for seen image reconstructions (Supplementary Fig. 13), suggesting that this semantic assistance compensates for the lack of low-level visual information in the imagery reconstruction.

Discussion

To reconstruct mental imagery from brain activity, we extended the previous image reconstruction method proposed by Shen et al. (2019) to a Bayesian estimation framework and introduced the assistance of semantic information (Fig. 1). While this previous reconstruction method was highly dependent on low-level visual information (i.e., low-layer VGG representations) decoded from the brain, our current proposed framework successfully produced meaningful reconstructions only using high-level visual information (Fig. 2). This advantage allows for the better use of such high-level visual information retained in the brain during imagery, thereby enabling mental image reconstruction (Fig. 3). Subsequent ablation analyses demonstrated that the two components introduced into our framework (Bayesian estimation and assistance of semantic information) were necessary for meaningful reconstructions (Figs. 6 and 7).

Our framework could reconstruct artificial shapes, even though the brain decoders were solely trained with fMRI responses to 1200 natural images (Figs. 2 and 3, Supplementary Figs. 4–7, 9, and 10). These results demonstrate that our reconstruction framework has a strong generalization ability for images in a new unknown domain, excluding the possibility of generating images by virtually selecting them from limited exemplars. As our spontaneous thoughts are not controlled or limited in daily life, such a strong generalization ability is potentially helpful for brain–machine interface applications in practical situations. In addition, the Bayesian nature of our framework would allow for better reconstruction by using a customized image prior when the domain of the image to be reconstructed is known in advance. According to a previous study (Schoenmakers et al., 2013), using a suitable prior improves the reconstruction of hand-written letter images from brain activity. Exploring this Bayesian advantage with a variety of image domains or sensory modalities other than vision would be an important challenge in the field of neural decoding.

We used the SGLD algorithm (Welling & Teh, 2011) to obtain samples (i.e., reconstructed images) from the Bayesian posterior distribution. In Bayesian estimation, sampling from the posterior distribution is often intractable. Thus, the use of Bayesian estimation for visual image reconstruction has been limited. A few previous studies have introduced Bayesian estimation into visual image reconstruction, which can be divided into two types: 1) cases in which the posterior distribution is analytically obtained, and 2) cases in which the maximum a posteriori probability estimate is approximately obtained by selecting it from a finite set of samples. For example, Schoenmakers et al. (2013) adopted the first approach and demonstrated that their Bayesian framework produced better reconstructions of hand-written letter images. In this case, a multivariate Gaussian distribution was used as the image prior. A similar concept has been applied to face image reconstruction in a subsequent study (Güçlütürk et al., 2017). However, given that the posterior distribution must be obtained analytically, this first approach cannot be combined with more general and flexible priors. In the second

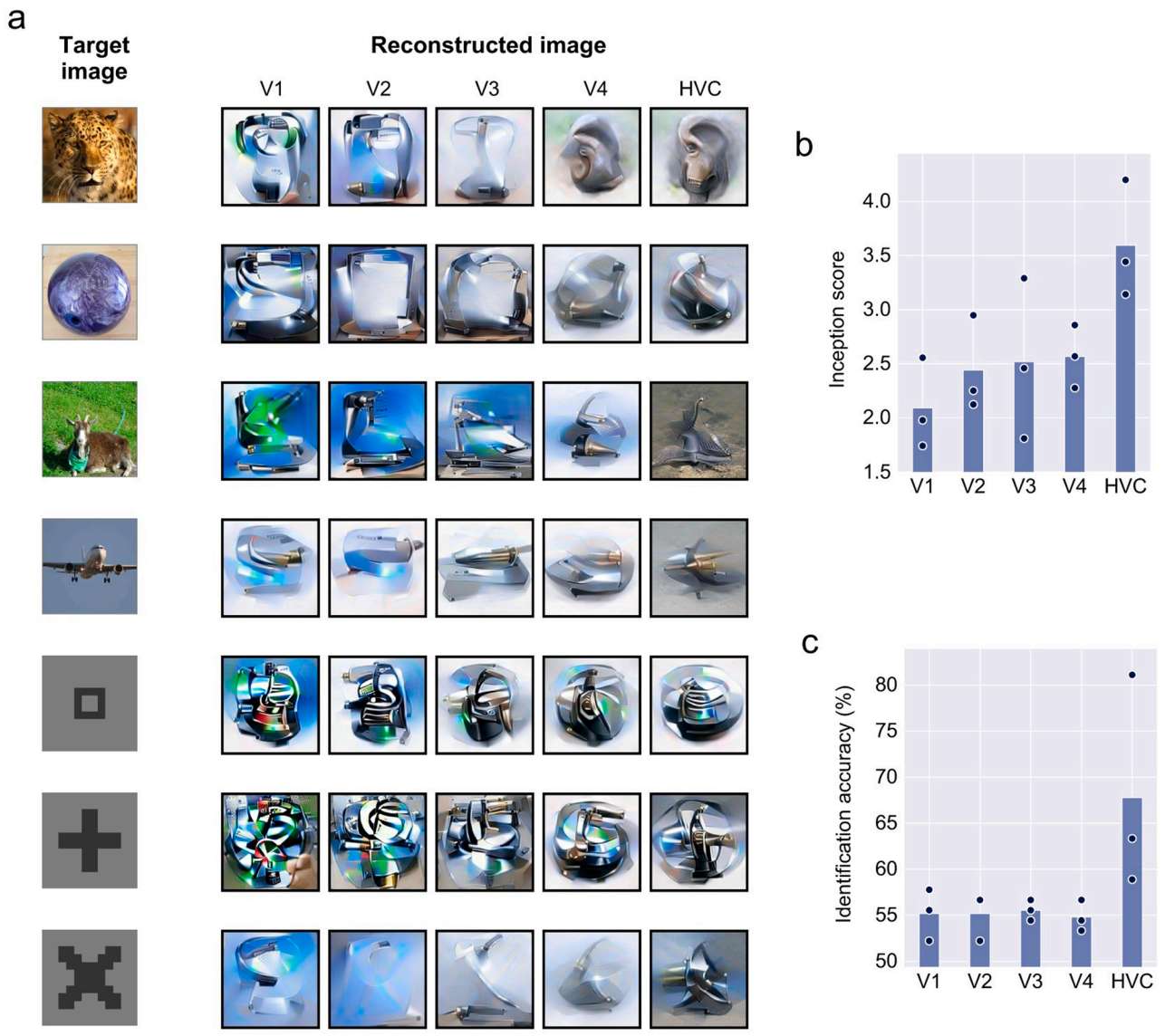

Fig. 5. Comparison of imagery reconstructions between input brain areas. (a) Imagery reconstructions from brain areas V1, V2, V3, V4, and the higher visual cortex (HVC). Images reconstructed via conv2–fc6 are shown. **(b)** Inception score. **(c)** Image identification accuracy.

approach, a set of images is independently prepared in advance, and the image with the highest posterior probability is treated as the reconstructed image (Naselaris et al., 2009; Qiao et al., 2020). Theoretically, this method would work if an infinite or sufficient number of samples have been prepared, but cannot reconstruct arbitrary images with limited exemplars. In our study, we introduced the SGLD algorithm into the reconstruction framework as an alternative approach and demonstrated that an image prior constructed with a pre-trained neural network improved the quality of reconstructions (Fig. 6). These results demonstrate the effectiveness of the SGLD algorithm for neural decoding.

While we employed the SGLD algorithm, which is one of the Bayesian sampling methods, exploring optimization algorithms suitable for our framework is an interesting direction. In the past several decades, a considerable array of optimization algorithms applicable to optimization problems with few or no assumptions on the objective function have been proposed (Babalola et al., 2020; Beheshti & Shamsuddin, 2013; S. Cheng et al., 2016; Kumar et al., 2023). Recently, several studies demonstrated that optimization algorithms inspired by natural phenomena are efficient for solving practical problems (Agushaka et al., 2022, 2023; Ezugwu et al., 2022; Hu et al., 2023; Laith et al., 2023; Zare et al., 2023). Those novel powerful nature-inspired algorithms have a lot of potential to improve our mental image reconstruction process.

The image generation process of our framework resembles that of a text-to-image generation method. In a popular algorithm for text-to-image generation, an image is generated by optimizing the latent vector of a pre-trained image generator model such that the output image matches the target text in the multi-modal embedding space provided by the CLIP model (Crowson et al., 2022). Thus, although our reconstruction framework was derived from Shen et al. (2019), it can be considered an extension of such a text-to-image generation algorithm for brain-to-image generation. Additionally, brain-to-text generation as a fusion of the above two types of algorithms would be an interesting future topic in the field of neural decoding (Huang et al., 2021; Tang et al., 2023).

While our reconstruction framework provides fundamental technology for brain–machine interfaces, it also serves as a tool for investigating the generation process of mental imagery. The comparison of input brain areas in the human visual hierarchy showed that the highest quality of imagery reconstruction was achieved with HVC (Fig. 5). These results are consistent with previous neuroimaging studies supporting the idea that HVC is recruited more than the lower visual areas during

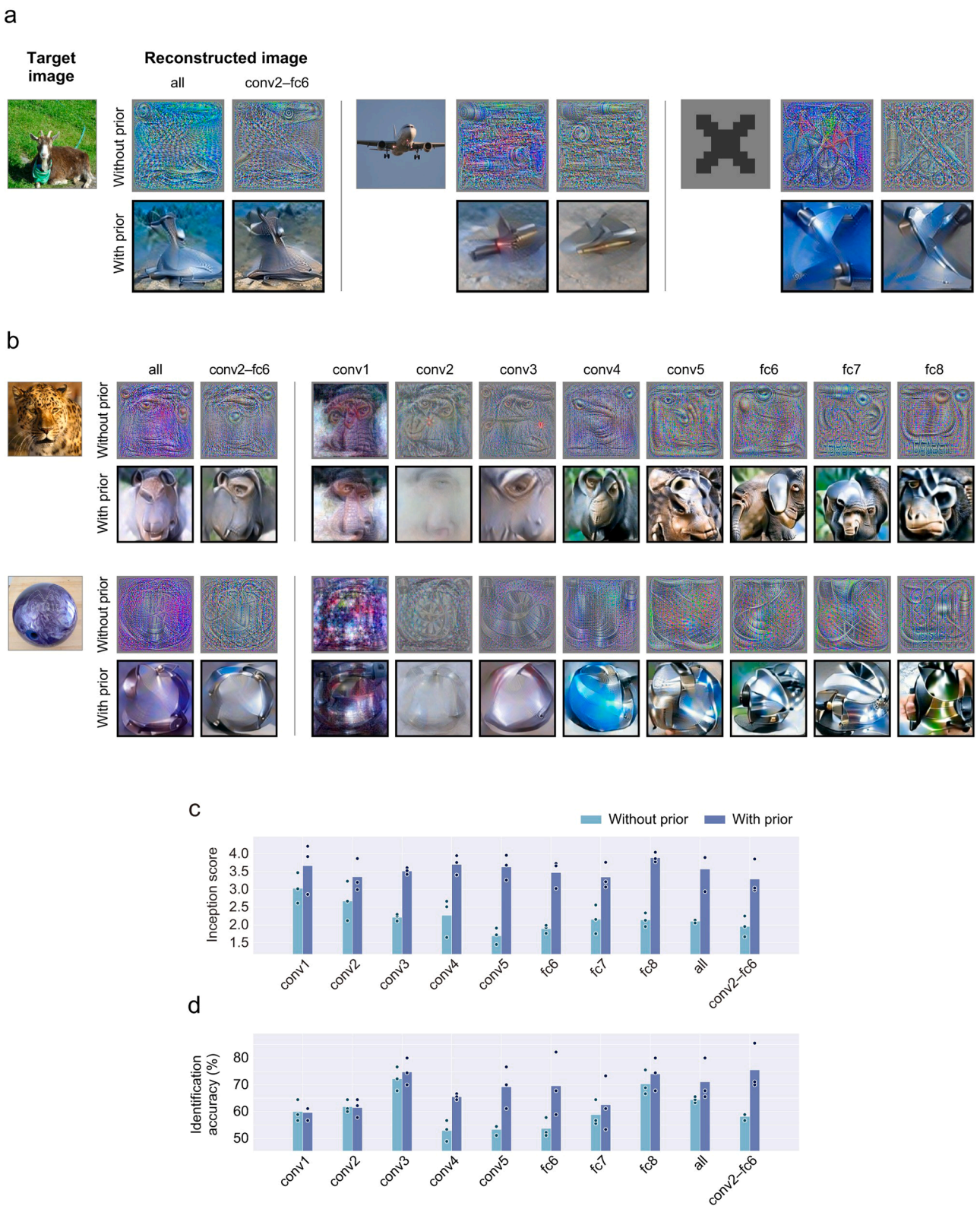

Fig. 6. Effect of the image prior. Imagery reconstruction results are compared between our full method and that without the image prior. The formats and target images are the same as those in Fig. 3.

imagery. Furthermore, we found that the line components in the imagery reconstructions of some artificial shapes were emphasized compared to those in the seen image reconstructions (Fig. 4). Although further investigation is required, this finding probably reflects the sharpening effect caused by the top-down process in the brain (Abdelhack & Kamitani, 2018). Therefore, our framework provides a novel approach for investigating hypotheses regarding mental imagery.

In line with recent attempts to reconstruct imagined speech (Moses et al., 2021; Tang et al., 2023), our proposed framework potentially paves the way for mind-reading technology, underscoring the necessity

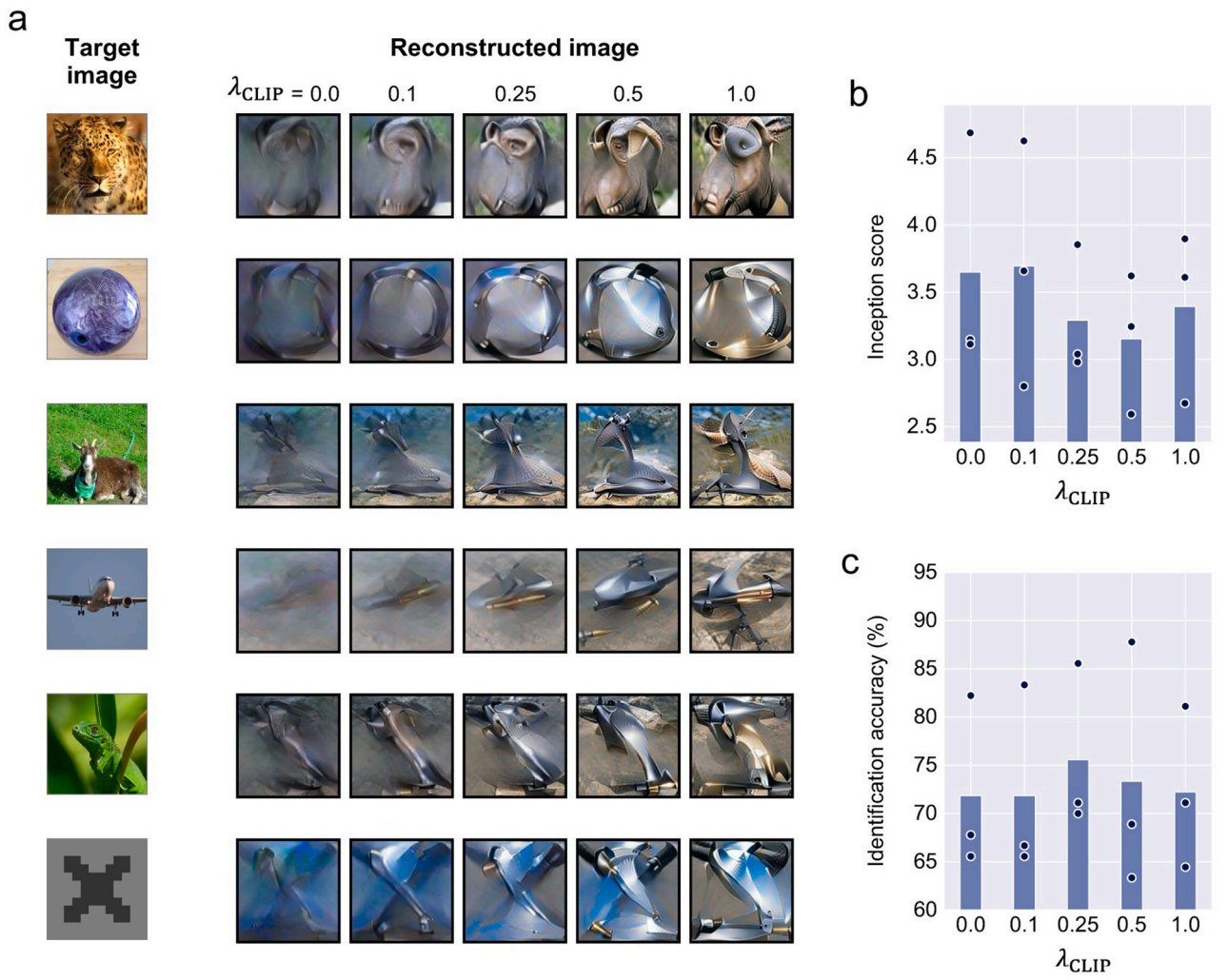

Fig. 7. Effect of the assistance of semantic information. (a) Imagery reconstructions with different strengths of semantic information assistance. The hyperparameter controlling the strength of semantic information assistance varied from 0.0 to 1.0. Images reconstructed via conv2–fc6 are shown. **(b)** Inception score. **(c)** Image identification accuracy.

of addressing ethical concerns associated with mental privacy (Rainey et al., 2020). To better understand the limitations and potential capabilities of our proposed framework, we present two critical research questions for future exploration: 1) Can brain decoders be trained without the subject's cooperation? 2) Can imagery spontaneously arising in the mind be accurately reconstructed? Regarding the first question, in our study, brain decoders were trained using fMRI responses to 1200 natural images and those decoders were trained and tested with the same subjects. While such within-subject decoding is relatively common, methods for constructing across-subject decoders have been recently proposed (Bilenko & Gallant, 2016; P.-H. (Cameron) Chen et al., 2015; Guntupalli et al., 2016; Haxby et al., 2011; Ho et al., 2023; Van Uden et al., 2018; Yamada et al., 2015). These techniques allow the construction of brain decoders for new subjects using minimal or no additional data, potentially leading to the development of zero-shot mind-reading technology. Regarding the second question, it is important to note that the fMRI data used in our study were measured while the subjects imagined target images specified by the experimenters (see Section 3.1; also see (Shen et al., 2019) for detailed information). Investigating the capability of our framework to decode spontaneous thoughts independent of the subject's intention remains a challenge, and it would offer critical insights into mental privacy concerns.

Conclusions

In this study, we have presented a machine learning method for visualizing subjective images in the human mind based on fMRI brain activity. While numerous previous studies provided machine learning methods to reconstruct visual stimuli from brain activity, the visualization of mental imagery had been left as a significant challenge. A few studies have reported successful visualization of mental imagery; however, their visualizable images were limited to specific domains such as faces, alphabetical letters, or geometric shapes. To transcend these limitations, we enhanced one of the previous visual image reconstruction methods through the integration of Bayesian estimation and semantic assistance. The experimental results demonstrated the capabilities of our proposed framework in reconstructing both natural images and artificial shapes that were imagined by human participants. The subsequent quantitative assessments also revealed that imagined images could be identified by our reconstructed images highly accurately compared to the chance accuracy. These promising results not only highlight the efficacy of our proposed framework but also suggest its potential as a unique tool for directly delving into the subjective contents of the brain, encompassing phenomena such as illusions, hallucinations, and dreams.

Code availability

The codes for our proposed reconstruction framework will be available at our GitHub repository (https://github.com/nkmjm/mental_img_recon) soon after the publication of the journal version of this article. All resources used to reproduce the results of Shen et al. (2019) are available from open repositories (https://github.com/KamitaniLab/DeepImageReconstruction, https://figshare.com/articles/dataset/Deep_Image_Reconstruction/7033577).

CRediT authorship contribution statement

NK and KM: designed the study. **NK and KM**: proposed the algorithms. **NK:** performed data analyses. **NK, SN, and KM:** interpreted the data. **NK, SN, and KM** wrote the manuscript.

Funding

This research was supported by MEXT Q-LEAP (JPMXS0120330644), JSPS KAKENHI (20K16465), JST ERATO (JPMJER1801), JST PRESTO (JPMJPR2128), and JST CREST (JPMJCR18A5 and JPMJCR22P3).

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

We used the fMRI data from Shen et al. (2019). These are available in an open data repository (https://figshare.com/articles/dataset/Deep_Image_Reconstruction/7033577).

Acknowledgments

The authors express their gratitude to Susumu Saito, Hidehiko Takahashi, Shuntaro Aoki, Fan Cheng, Yukiyasu Kamitani, and members of Kamitani laboratory at Kyoto university.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neunet.2023.11.024.

References

Abdelhack, M., & Kamitani, Y. (2018). Sharpening of hierarchical visual feature representations of blurred images. *eNeuro, 5*(3). https://doi.org/10.1523/ENEURO.0443-17.2018. ENEURO.0443-17.2018.

Agushaka, J. O., Ezugwu, A. E., & Abualigah, L. (2022). Dwarf mongoose optimization algorithm. *Computer Methods in Applied Mechanics and Engineering, 391*, Article 114570. https://doi.org/10.1016/j.cma.2022.114570

Agushaka, J. O., Ezugwu, A. E., & Abualigah, L. (2023). Gazelle optimization algorithm: A novel nature-inspired metaheuristic optimizer. *Neural Computing and Applications, 35*(5), 4099–4131. https://doi.org/10.1007/s00521-022-07854-6

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & de Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology: CB, 23*(15), 1427–1431. https://doi.org/10.1016/j.cub.2013.05.065

Babalola, A. E., Ojokoh, B. A., & Odili, J. B. (2020). A review of population-based optimization algorithms. In *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)* (pp. 1–7). https://doi.org/10.1109/ICMCECS47690.2020.240856

Beheshti, Z., & Shamsuddin, S. M. H. (2013). A review of population-based meta-heuristic algorithm. *International Journal of Advances in Soft Computing and Its Applications, 5*(1).

Beliy, R., Gaziv, G., Hoogi, A., Strappini, F., Golan, T., & Irani, M. (2019). From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. *Advances in Neural Information Processing Systems*, 32.

Bilenko, N. Y., & Gallant, J. L. (2016). Pyrcca: Regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. *Frontiers in Neuroinformatics, 10*, 49. https://doi.org/10.3389/fninf.2016.00049

Brock, A., Donahue, J., & Simonyan, K. (2019). *Large scale GAN training for high fidelity natural image synthesis* (arXiv:1809.11096). arXiv. http://arxiv.org/abs/1809.11096.

Chang, C. H., Zehra, S., Nestor, A., & Lee, A. C. H. (2023). Using image reconstruction to investigate face perception in amnesia. *Neuropsychologia, 185*, Article 108573. https://doi.org/10.1016/j.neuropsychologia.2023.108573

(Cameron) Chen, P. H, Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). A reduced-dimension fMRI shared response model. In *Advances in Neural Information Processing Systems, 28* https://papers.nips.cc/paper_files/paper/2015/hash/b3967a0e938dc2a6340e258630febd5a-Abstract.html.

Chen, Z., Qing, J., Xiang, T., Yue, W.L., & Zhou, J.H. (2023). *Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding* (arXiv:2211.06956). arXiv. http://arxiv.org/abs/2211.06956.

Cheng, S., Liu, B., Ting, T. O., Qin, Q., Shi, Y., & Huang, K. (2016). Survey on data science with population-based algorithms. *Big Data Analytics, 1*(1), 3. https://doi.org/10.1186/s41044-016-0003-3

Cheng, F., Horikawa, T., Majima, K., Tanaka, M., Abdelhack, M., Aoki, S. C., et al. (2023). Reconstructing visual illusory experiences from human brain activity [Preprint] *Neuroscience*. https://doi.org/10.1101/2023.06.15.545037.

Cichy, R. M., Heinzle, J., & Haynes, J. D. (2012). Imagery and perception share cortical representations of content and location. In *Cerebral Cortex (New York, N.Y.: 1991), 22* pp. 372–380). https://doi.org/10.1093/cercor/bhr106

Cowen, A. S., Chun, M. M., & Kuhl, B. A. (2014). Neural portraits of perception: Reconstructing face images from evoked brain activity. *NeuroImage, 94*, 12–22. https://doi.org/10.1016/j.neuroimage.2014.03.018

Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L. et al. (2022). VQGAN-CLIP: Open domain image generation and editing with natural language guidance. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer vision – eccv 2022* (Vol. 13697, pp. 88–105). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19836-6_6.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, Kai, & Fei-Fei, Li (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

Dhariwal, P., & Nichol, A. (2021). *Diffusion models beat GANs on image synthesis* (arXiv:2105.05233). arXiv. http://arxiv.org/abs/2105.05233.

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. J. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences, 23*(5), 423–434. https://doi.org/10.1016/j.tics.2019.02.004

Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.48550/ARXIV.2012.09841

Ezugwu, A. E., Agushaka, J. O., Abualigah, L., Mirjalili, S., & Gandomi, A. H. (2022). Prairie dog optimization algorithm. *Neural Computing and Applications, 34*(22), 20017–20065. https://doi.org/10.1007/s00521-022-07530-9

Fang, T., Qi, Y., & Pan, G. (2020). Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems, 33*, 13038–13048.

Fujiwara, Y., Miyawaki, Y., & Kamitani, Y. (2013). Modular encoding and decoding models derived from Bayesian canonical correlation analysis. *Neural Computation, 25*(4), 979–1005. https://doi.org/10.1162/NECO_a_00423

Fukuma, R., Yanagisawa, T., Nishimoto, S., Sugano, H., Tamura, K., Yamamoto, S., et al. (2022). Voluntary control of semantic neural representations by imagery with conflicting visual stimulation. *Communications Biology, 5*(1), 214. https://doi.org/10.1038/s42003-022-03137-x

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., & van Gerven, M. A. J. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in Neural Information Processing Systems, 30*. https://proceedings.neurips.cc/paper/2017/hash/efdf562ce2fb0ad460fd8e9d33e57f57-Abstract.html.

Gaziv, G., Beliy, R., Granot, N., Hoogi, A., Strappini, F., Golan, T., et al. (2022). Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage, 254*, Article 119121. https://doi.org/10.1016/j.neuroimage.2022.119121

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., & Haxby, J. V. (2016). A model of representational spaces in human cortex. In *Cerebral Cortex (New York, N.Y.: 1991), 26* pp. 2919–2934). https://doi.org/10.1093/cercor/bhw068

Han, K., Wen, H., Shi, J., Lu, K. H., Zhang, Y., Fu, D., et al. (2019). Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage, 198*, 125–136. https://doi.org/10.1016/j.neuroimage.2019.05.039

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature, 458*(7238), 632–635. https://doi.org/10.1038/nature07832

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., et al. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron, 72*(2), 404–416. https://doi.org/10.1016/j.neuron.2011.08.026

Ho, J. K., Horikawa, T., Majima, K., Cheng, F., & Kamitani, Y. (2023). Inter-individual deep image reconstruction via hierarchical neural code conversion. *NeuroImage, 271*, Article 120007. https://doi.org/10.1016/j.neuroimage.2023.120007

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications, 8*, 15037. https://doi.org/10.1038/ncomms15037

Horikawa, T., & Kamitani, Y. (2022). Attention modulates neural representation to render reconstructions according to subjective appearance. *Communications Biology, 5*(1), 34. https://doi.org/10.1038/s42003-021-02975-5

Hu, G., Zheng, Y., Abualigah, L., & Hussien, A. G. (2023). DETDO: An adaptive hybrid dandelion optimizer for engineering optimization. *Advanced Engineering Informatics, 57*, Article 102004. https://doi.org/10.1016/j.aei.2023.102004

Huang, W., Yan, H., Cheng, K., Wang, C., Li, J., Wang, Y., et al. (2021). A neural decoding algorithm that generates language from visual activity evoked by natural images. *Neural Networks: The Official Journal of the International Neural Network Society, 144*, 90–100. https://doi.org/10.1016/j.neunet.2021.08.006

Jafari-Khouzani, K., & Soltanian-Zadeh, H. (2005). Radon transform orientation estimation for rotation invariant texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(6), 1004–1008. https://doi.org/10.1109/TPAMI.2005.126

Kay, K. N., & Gallant, J. L. (2009). I can see what you see. *Nature Neuroscience, 12*(3), 245. https://doi.org/10.1038/nn0309-245

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452*(7185), 352–355. https://doi.org/10.1038/nature06713

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*. https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Kumar, A., Nadeem, M., & Banka, H. (2023). Nature inspired optimization algorithms: A comprehensive overview. *Evolving Systems, 14*(1), 141–156. https://doi.org/10.1007/s12530-022-09432-6

Laith, A., Serdar, E., Davut, I., & Abu, Z. R. (2023). Modified elite opposition-based artificial hummingbird algorithm for designing FOPID controlled cruise control system. *Intelligent automation & soft computing*.

Lee, H., & Kuhl, B. A. (2016). Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 36*(22), 6069–6082. https://doi.org/10.1523/JNEUROSCI.4286-15.2016

Lee, S. H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. *NeuroImage, 59*(4), 4064–4073. https://doi.org/10.1016/j.neuroimage.2011.10.055

Lu, Y., Du, C., Wang, D., & He, H. (2023). *MindDiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion* (arXiv:2303.14139). arXiv. http://arxiv.org/abs/2303.14139.

Majima, K., Sukhanov, P., Horikawa, T., & Kamitani, Y. (2017). Position information encoded by population activity in hierarchical visual areas. *eNeuro, 4*(2). https://doi.org/10.1523/ENEURO.0268-16.2017. ENEURO.0268-16.2017.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H. C., et al. (2008). Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders. *Neuron, 60*(5), 915–929. https://doi.org/10.1016/j.neuron.2008.11.004

Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., et al. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *The New England Journal of Medicine, 385*(3), 217–227. https://doi.org/10.1056/NEJMoa2027540

Mozafari, M., Reddy, L., & VanRullen, R. (2020). Reconstructing natural scenes from fMRI patterns using BigBiGAN. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). https://doi.org/10.1109/IJCNN48605.2020.9206960

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron, 63*(6), 902–915. https://doi.org/10.1016/j.neuron.2009.09.006

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage, 105*, 215–228. https://doi.org/10.1016/j.neuroimage.2014.10.018

Nestor, A., Lee, A. C. H., Plaut, D. C., & Behrmann, M. (2020). The face of image reconstruction: Progress, pitfalls, prospects. *Trends in Cognitive Sciences, 24*(9), 747–759. https://doi.org/10.1016/j.tics.2020.06.006

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology: CB, 21*(19), 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031

Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience, 24*(9), Article 103013. https://doi.org/10.1016/j.isci.2021.103013

Oord, A.van den, Vinyals, O., & Kavukcuoglu, K. (2018). *Neural discrete representation learning* (arXiv:1711.00937). arXiv. http://arxiv.org/abs/1711.00937.

Ozcelik, F., & VanRullen, R. (2023). *Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion* (arXiv:2303.05334). arXiv. http://arxiv.org/abs/2303.05334.

Qiao, K., Zhang, C., Wang, L., Chen, J., Zeng, L., Tong, L., et al. (2018). Accurate reconstruction of image stimuli from human functional magnetic resonance imaging based on the decoding model with capsule network architecture. *Frontiers in Neuroinformatics, 12*, 62. https://doi.org/10.3389/fninf.2018.00062

Qiao, K., Chen, J., Wang, L., Zhang, C., Tong, L., & Yan, B. (2020). BigGAN-based Bayesian reconstruction of natural images from human brain activity. *Neuroscience, 444*, 92–105. https://doi.org/10.1016/j.neuroscience.2020.07.040

Radford, A., Metz, L., & Chintala, S. (2016). *Unsupervised representation learning with deep convolutional generative adversarial networks* (arXiv:1511.06434). arXiv. http://arxiv.org/abs/1511.06434.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S. et al. (2021). *Learning transferable visual models from natural language supervision*. https://doi.org/10.48550/ARXIV.2103.00020.

Rainey, S., Martin, S., Christen, A., Mégevand, P., & Fourneret, E. (2020). Brain recording, mind-reading, and neurotechnology: Ethical issues from consumer devices to brain-based speech decoding. *Science and Engineering Ethics, 26*(4), 2295–2311. https://doi.org/10.1007/s11948-020-00218-0

Rakhimberdina, Z., Jodelet, Q., Liu, X., & Murata, T. (2021). Natural image reconstruction from fMRI using deep learning: A survey. *Frontiers in Neuroscience, 15*, Article 795488. https://doi.org/10.3389/fnins.2021.795488

Razavi, A., Oord, A.van den, & Vinyals, O. (2019). *Generating diverse high-fidelity images with VQ-VAE-2* (arXiv:1906.00446). arXiv. http://arxiv.org/abs/1906.00446.

Reddy, L., Tsuchiya, N., & Serre, T. (2010). Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage, 50*(2), 818–825. https://doi.org/10.1016/j.neuroimage.2009.11.084

Ren, Z., Li, J., Xue, X., Li, X., Yang, F., Jiao, Z., et al. (2021). Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage, 228*, Article 117602. https://doi.org/10.1016/j.neuroimage.2020.117602

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-resolution image synthesis with latent diffusion models* (arXiv:2112.10752). arXiv. http://arxiv.org/abs/2112.10752.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). *Improved techniques for training GANs* (arXiv:1606.03498). arXiv. http://arxiv.org/abs/1606.03498.

Satake, E., Majima, K., Aoki, S. C., & Kamitani, Y. (2018). Sparse ordinal logistic regression and its application to brain decoding. *Frontiers in Neuroinformatics, 12*, 51. https://doi.org/10.3389/fninf.2018.00051

Schoenmakers, S., Barth, M., Heskes, T., & van Gerven, M. (2013). Linear reconstruction of perceived images from human brain activity. *NeuroImage, 83*, 951–961. https://doi.org/10.1016/j.neuroimage.2013.07.043

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & van Gerven, M. A. J. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage, 181*, 775–785. https://doi.org/10.1016/j.neuroimage.2018.07.043

Senden, M., Emmerling, T. C., van Hoof, R., Frost, M. A., & Goebel, R. (2019). Reconstructing imagined letters from early visual cortex reveals tight topographic correspondence between visual mental imagery and perception. *Brain Structure & Function, 224*(3), 1167–1183. https://doi.org/10.1007/s00429-019-01828-6

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLOS Computational Biology, 15*(1), Article e1006633. https://doi.org/10.1371/journal.pcbi.1006633

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. https://doi.org/10.48550/ARXIV.1409.1556.

Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., & Poole, B. (2021). *Score-based generative modeling through stochastic differential equations* (arXiv:2011.13456). arXiv. http://arxiv.org/abs/2011.13456.

St-Yves, G., & Naselaris, T. (2018). Generative adversarial networks conditioned on brain activity reconstruct seen images. In *Conference Proceedings. IEEE International Conference on Systems, Man, and Cybernetics, 2018* (pp. 1054–1061). https://doi.org/10.1109/SMC.2018.00187

Stokes, M., Thompson, R., Cusack, R., & Duncan, J. (2009). Top-down activation of shape-specific population codes in visual cortex during mental imagery. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 29*(5), 1565–1572. https://doi.org/10.1523/JNEUROSCI.4657-08.2009

Takagi, Y., & Nishimoto, S. (2023). *High-resolution image reconstruction with latent diffusion models from human brain activity*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 14453–14463). https://doi.org/10.1109/CVPR52729.2023.01389

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience, 26*(5), 858–866. https://doi.org/10.1038/s41593-023-01304-9

Van Uden, C. E., Nastase, S. A., Connolly, A. C., Feilong, M., Hansen, I., Gobbini, M. I., et al. (2018). Modeling semantic encoding in a common neural representational space. *Frontiers in Neuroscience, 12*, 437. https://doi.org/10.3389/fnins.2018.00437

VanRullen, R., & Reddy, L. (2019). Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications Biology, 2*, 193. https://doi.org/10.1038/s42003-019-0438-y

Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *, 8. Proceedings of the 28th International Conference on International Conference on Machine Learning*.

Xing, Y., Ledgeway, T., McGraw, P. V., & Schluppeck, D. (2013). Decoding working memory of stimulus contrast in early visual cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 33*(25), 10301–10311. https://doi.org/10.1523/JNEUROSCI.3754-12.2013

Yamada, K., Miyawaki, Y., & Kamitani, Y. (2015). Inter-subject neural code converter for visual image representation. *NeuroImage, 113*, 289–297. https://doi.org/10.1016/j.neuroimage.2015.03.059

Zare, M., Ghasemi, M., Zahedi, A., Golalipour, K., Mohammadi, S. K., Mirjalili, S., et al. (2023). A global best-guided firefly algorithm for engineering problems. *Journal of Bionic Engineering, 20*(5), 2359–2388. https://doi.org/10.1007/s42235-023-00386-2