Decoding and Reconstructing Visual Experience from Brain Activity with Generative Latent Representations

Motokazu Umehara[1], Misato Tanaka[1,2], Yoshihiro Nagano[1,2], Yukiyasu Kamitani[1,2,3]
[1]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[2]ATR Computational Neuroscience Laboratories, Kyoto, Japan
[3]Guardian Robot Project, RIKEN, Kyoto, Japan
umehara.motokazu.85e@st.kyoto-u.ac.jp, tanaka.misato.4f@kyoto-u.ac.jp
nagano@i.kyoto-u.ac.jp, kamitani@i.kyoto-u.ac.jp

Abstract

The brain's bidirectional processing is thought to parallel the functions of recognition (bottom-up) and generative (top-down) models in AI. While prior work has established a hierarchical correspondence between the visual cortex and recognition models, whether generative models exhibit a similar neural alignment remains an open question. To investigate this, we present a unified framework that decodes fMRI signals into the latent representations of both model classes for visual reconstruction. Our analysis of perception data revealed that the generative model achieved decoding and reconstruction performance comparable to the recognition model, though its hierarchical correspondence with the visual cortex was weaker and followed a different trend. An application to imagery data showed low decoding accuracies but a different pattern from perception. Our work contributes a comparative pipeline for studying generative representations and provides a framework for future investigations into the brain's bidirectional processing architecture.

1 Introduction

The brain processes information through bottom-up and top-down pathways. For example, when recognizing objects in an image, the brain integrates bottom-up information based on visual stimuli with top-down information derived from prior knowledge [1–3]. These processes are thought to correspond, respectively, to a recognition model that takes an image as input and a generative model that produces an image as output. By comparing these two model classes, we can better characterize how the brain represents information along both directions of processing, thereby providing insight into the bidirectional nature of neural computation.

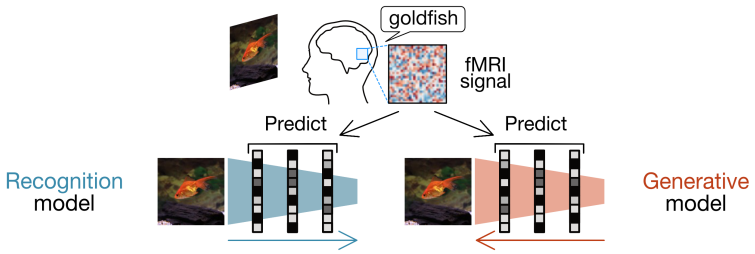

Figure 1: Overview of the study pipeline. The hierarchical alignment of generative and recognition models with the brain is evaluated by decoding the models' latent representations from fMRI data.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CogInterp.

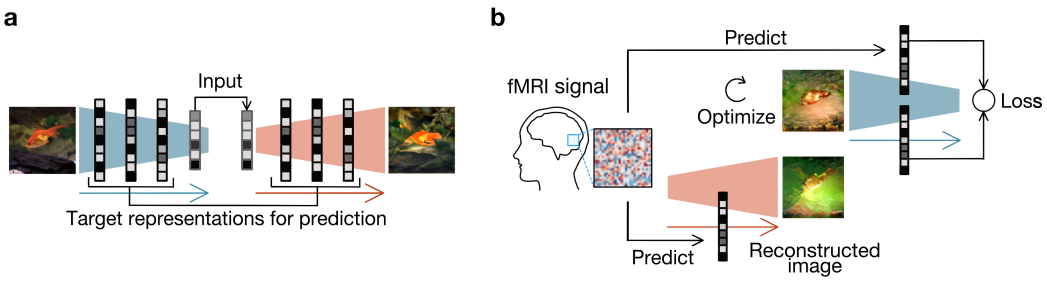

Figure 2: (a) The relationship between the recognition and generative models used in this work, along with the target representations for prediction. (b) Reconstruction pipeline from predicted latent representations of the recognition and generative models.

With the advancement of deep learning, NeuroAI research on the correspondence between models' representations and brain activity has gained attention [4–9]. In particular, the hierarchical correspondence between layers of recognition models and the human visual system has been widely studied: their features can be predicted from brain activity [10] and used to reconstruct images [11, 12]. In contrast, the hierarchical correspondence of generative models with the brain is less understood. Although latent seeds of generative models have been shown to be decoded from brain activity to reconstruct perceived images [13, 14] and illusory images [15], how their hierarchical representations map onto the brain's visual hierarchy remains unclear (See the details in Appendix A).

To address this question, we developed a method to decode the representations of the recognition and generative models from fMRI data and compare them (Figure 1). We found that the generative model exhibited decoding and reconstruction performance comparable to the recognition model. However, the hierarchical correspondence of the generative model was not as strong as that of the recognition model, especially in layers processing low-level image features. An application to imagery data, a more top-down task, showed a different pattern from perception, although overall accuracy was low and results remained inconclusive. Our work provides a comparative pipeline for studying generative representations and a framework for future investigations of bidirectional processing in the brain.

2 Methods

We first trained decoders to predict the hierarchical representations of the recognition and generative models from fMRI signals, following the methods of Shen et al. [11]. We used the representations of the following DNN models as targets (Figure 2a): the recognition model was a variant of AlexNet [16], and the generative model was pretrained to recover images from single-layer features of this recognition model [17]. These models are approximately symmetric, enabling a fair comparison. To train the decoders, we used fMRI data from three subjects provided by Shen et al. [18], collected during visual perception. We trained decoders using both the entire visual cortex and individual visual areas. Further details of the dataset, models, and training are provided in Appendix B.1–B.3.

Using the trained decoders, we then predicted the latent representations of the recognition and generative models and reconstructed images from them. From the predicted representations, reconstructed images were generated for each layer using the following procedures (Figure 2b): for the recognition model, we optimized the image to minimize the loss between its representation and the predicted representation; for the generative model, we directly fed the predicted representation into the model to generate the image. In the test phase, we used the held-out fMRI data during visual perception and mental imagery. Our main analyses focused on perception, and imagery data were used as an additional application. Further details of the dataset and reconstruction procedures are provided in Appendix B.1 and B.4. Based on these predicted representations and reconstructed images, we conducted subsequent analyses described in the Results section.

3 Results

We first examined perception data from the whole visual cortex to confirm whether representations of the generative model could be decoded and used for reconstruction. We evaluated decoding

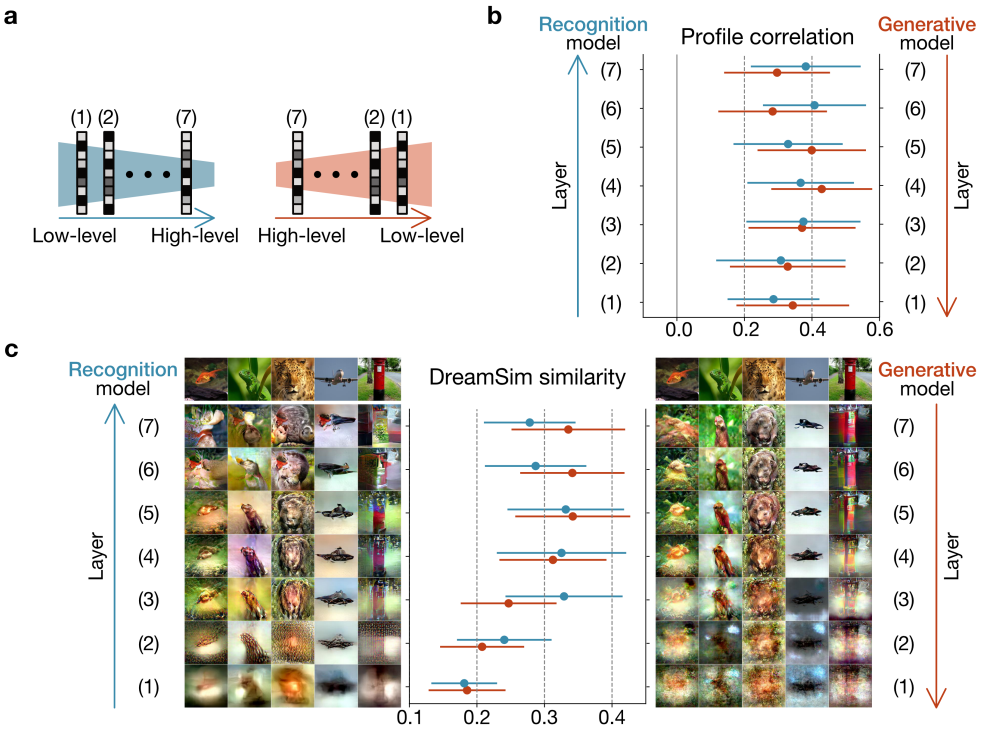

Figure 3: (a) Correspondence between models and layer numbers, increasing from low- to high-level features. (b) Profile correlations of decoded representations from perception data in the whole visual cortex (mean ± SD across all units pooled from three subjects). (c) Reconstructed images (from a representative subject) and their DreamSim similarity evaluation using the same perception data (mean ± SD across all stimuli pooled from three subjects). In all panels, blue denotes the recognition model, and red denotes the generative model.

and reconstruction accuracies at each layer (Figure 3a). Decoding accuracy was evaluated using *profile correlation*, which measures the Pearson correlation between predicted and target values across stimuli for each unit. Profile correlation values showed that representations of the generative model could be predicted at a level similar to the recognition model and comparable to prior studies [10, 11], although layers processing high-level image features showed slightly lower values only for the generative model (Figure 3b). Both models showed high correlations in mid-level layers.

Reconstructed images from the generative model captured the overall shape of the targets across most layers, similar to those from the recognition model (Figure 3c). We evaluated these reconstructed images using *DreamSim similarity* [19], which measures perceptual similarity between reconstructed and target images based on DNN features. DreamSim similarity values also indicated that decoded generative model representations could reconstruct images comparable to the recognition model. Similar to profile correlation, DreamSim similarity varied across layers: both models achieved the highest values in mid-level layers, while low-level layers yielded lower values. High-level layers showed relatively lower values for the recognition model. Overall, representations of the generative model could be decoded and used for reconstruction comparable to the recognition model.

Next, we used data from each visual area, which processes information at different hierarchical levels, to decode representations of the recognition and generative models and examined their hierarchical correspondence with the brain. We plotted the proportion of units best predicted by each visual area across model layers and calculated the *BH score* proposed by Nonaka et al. [20], which evaluates the hierarchical correspondence between DNN models and the brain. The generative model showed weaker correspondence with the brain than the recognition model, although it still exhibited alignment (Figure 4a). For high-level layers, representations of both models were better predicted in higher visual areas, though this trend was slightly weaker for the generative model. In contrast, for low-level layers, clear differences were observed: representations of the recognition model were better predicted in lower areas, whereas those of the generative model showed little variation across visual areas.

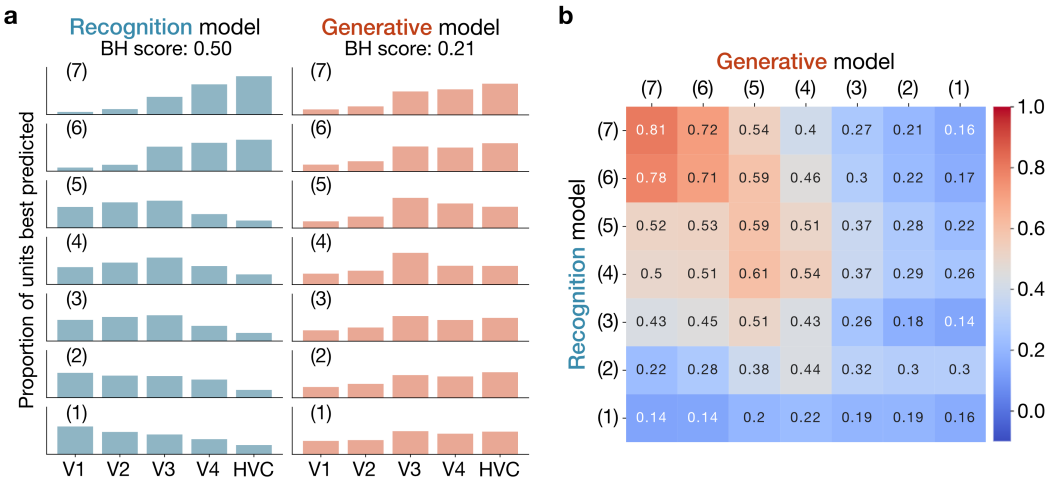

Figure 4: (a) BH scores and the proportion of units best predicted by each visual area across model layers, computed from the perception data (pooled across three subjects). Visual areas are shown in hierarchical order (V1-V4 and HVC). (b) Similarity matrix showing representational similarity between recognition and generative model layers, computed directly from the model representations without brain data.

To further investigate this difference, we visualized the representational similarity between recognition and generative models' layers, computed directly from their representations without brain data, using Representational Similarity Analysis (RSA) [21, 22]. We found that the two models were similar in high-level feature layers, but their similarity was lower in low-level layers (Figure 4b). These results indicate that, in terms of hierarchical correspondence, brain activity during perception aligns more closely with the representations of the recognition model than with those of the generative model.

Finally, we also applied these analyses to imagery data. Overall, decoding and reconstruction accuracies were very low, without substantial differences across models or layers. Hierarchical correspondence with the brain was weaker than in perception data for both models, and the recognition model showed a different trend. Detailed imagery results are provided in Appendix C.

4 Discussion

In this study, we developed a method to decode latent representations of the recognition and generative models from fMRI signals and reconstruct images from them. Our results from the whole visual cortex showed that the generative model achieved decoding and reconstruction performance comparable to the recognition model. In decoding, high-level layers performed better for the recognition model, possibly because these layers are task-optimized and contain information more consistent with brain activity. In reconstruction, the performance of the recognition model decreased in high-level layers, likely because the optimization-based reconstruction process is less influenced by layers farther from the image input. These differences suggest differences in brain correspondence between the two models, while also indicating the need to improve comparative methods.

Our results from each visual area indicated that the generative model showed a weaker hierarchical correspondence with the brain than the recognition model. In particular, for layers processing low-level image features, the generative model showed little variation across visual areas. This may be because its input originates from higher-level features, causing all layers to contain high-level information to some extent. When applied to imagery, a more top-down task, the recognition model's correspondence with the brain was particularly weakened, although decoding and reconstruction accuracies were generally low. While the present work is limited by factors such as the small number of participants and the use of pretrained models that were not strictly symmetric, focusing on these differences in brain correspondence between the two models may also provide insights into the brain's bidirectional information processing. Our work contributes a comparative pipeline for studying generative representations and offers a framework for future investigations into the brain's bidirectional processing architecture.

Acknowledgements

We would like to thank the members of our laboratory for their valuable feedback, helpful discussions, and continuous support throughout the course of this work. This work was supported by Japan Society for the Promotion of Science (JSPS: KAKENHI grants JP25H00450 to Y.K.) and Guardian Robot Project, RIKEN.

References

[1] Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis. 2003 Jul;20(7):1434–1448. doi: 10.1364/JOSAA.20.001434.

[2] Gilbert CD, Sigman M. Brain states: top-down influences in sensory processing. Neuron. 2007 Jun;54(5):677–696. doi: 10.1016/j.neuron.2007.05.019.

[3] Friston K. The free-energy principle: a unified brain theory? Nat Rev Neurosci. 2010 Feb;11(2):127–138. doi: 10.1038/nrn2787.

[4] Kriegeskorte N, Douglas PK. Cognitive computational neuroscience. Nat Neurosci. 2018 Sep;21(9):1148–1160. doi: 10.1038/s41593-018-0210-5.

[5] Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A. 2014 May;111(23):8619–8624. doi: 10.1073/pnas.1403112111.

[6] Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol. 2014 Nov;10(11):1–29. doi: 10.1371/journal.pcbi.1003915.

[7] Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, et al. Brain-Score: which artificial neural network for object recognition is most brain-like? bioRxiv [Preprint]. 2020 Jan 2 [posted 2020 Jan 2; cited 2025 Oct 15]: [9 p.]. Available from: `https://www.biorxiv.org/content/early/2020/01/02/407007` doi: 10.1101/407007.

[8] Ponce CR, Xiao W, Schade PF, Hartmann TS, Kreiman G, Livingstone MS. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. Cell. 2019 Sep;177(4):999-1009.e10. doi: 10.1016/j.cell.2019.04.005.

[9] Murty NAR, Bashivan P, Abate A, DiCarlo JJ, Kanwisher N. Computational models of category-selective brain regions enable high-throughput tests of selectivity. Nat Commun. 2021 Sep;12(1):5540. doi: 10.1038/s41467-021-25409-6.

[10] Horikawa T, Kamitani Y. Generic decoding of seen and imagined objects using hierarchical visual features. Nat Commun. 2017 May;8(1):15037. doi: 10.1038/ncomms15037.

[11] Shen G, Horikawa T, Majima K, Kamitani Y. Deep image reconstruction from human brain activity. PLoS Comput Biol. 2019 Jan;15(1):1–23. doi: 10.1371/journal.pcbi.1006633.

[12] Horikawa T, Kamitani Y. Attention modulates neural representation to render reconstructions according to subjective appearance. Commun Biol. 2022 Jan;5(1):34. doi: 10.1038/s42003-021-02975-5.

[13] Seeliger K, Güçlü U, Ambrogioni L, Güçlütürk Y, van Gerven MAJ. Generative adversarial networks for reconstructing natural images from brain activity. Neuroimage. 2018 Aug;181:775–785. doi: 10.1016/j.neuroimage.2018.07.043.

[14] Ozcelik F, VanRullen R. Natural scene reconstruction from fMRI signals using generative latent diffusion. Sci Rep. 2023 Sep;13(1):15666. doi: 10.1038/s41598-023-42891-8.

[15] Cheng FL, Horikawa T, Majima K, Tanaka M, Abdelhack M, Aoki SC, et al. Reconstructing visual illusory experiences from human brain activity. Sci Adv. 2023 Nov;9(46):eadj3906. doi: 10.1126/sciadv.adj3906.

[16] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1097–1105. Available from: `https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[17] Dosovitskiy A, Brox T. Generating images with perceptual similarity metrics based on deep networks. Adv Neural Inf Process Syst. 2016;29:658–666. Available from: `https://proceedings.neurips.cc/paper_files/paper/2016/file/371bce7dc83817b7893bcdeed13799b5-Paper.pdf`.

[18] Shen G, Horikawa T, Majima K, Kamitani Y. Deep image reconstruction from human brain activity; 2019 [cited 2025 Oct 15]. Database: figshare [Internet]. Available from: https://figshare.com/articles/dataset/Deep_Image_Reconstruction/7033577.

[19] Fu S, Tamir N, Sundaram S, Chai L, Zhang R, Dekel T, et al. DreamSim: learning new dimensions of human visual similarity using synthetic data. Adv Neural Inf Process Syst. 2023;36:50742–50768. Available from: https://proceedings.neurips.cc/paper_files/paper/2023/file/9f09f316a3eaf59d9ced5ffaefe97e0f-Paper-Conference.pdf.

[20] Nonaka S, Majima K, Aoki SC, Kamitani Y. Brain hierarchy score: which deep neural networks are hierarchically brain-like? iScience. 2021 Sep;24(9):103013. doi: 10.1016/j.isci.2021.103013.

[21] Kriegeskorte N, Mur M, Bandettini PA. Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci. 2008 Nov;2:4. doi: 10.3389/neuro.06.004.2008.

[22] Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. A toolbox for representational similarity analysis. PLoS Comput Biol. 2014 Apr;10(4):1–11. doi: 10.1371/journal.pcbi.1003553.

Appendix

A Related Work

Neural alignment between DNNs and the brain. The advancement of deep learning accelerated NeuroAI research on the correspondence between model and brain representations. Yamins et al. [5] showed that task-optimized DNN features predicted visual cortex responses. Khaligh-Razavi and Kriegeskorte [6] used representational similarity analysis and found that supervised DNNs align more closely with human brain activity than unsupervised models. Schrimpf et al. [7] proposed a quantitative benchmark, Brain-Score, to evaluate DNNs in terms of their neural similarity to the brain. Ponce et al. [8] and Murty et al. [9] developed techniques to generate images predicted to strongly activate specific brain regions based on the relationship between DNN representations and neural responses.

Hierarchical correspondence between DNNs and the brain. In the context of NeuroAI research, previous studies have further investigated the hierarchical correspondence between DNNs and the brain. Horikawa and Kamitani [10] decoded latent representations of recognition models from fMRI activity and demonstrated correspondence between the DNN layer hierarchy and the hierarchy of visual areas in terms of decoding accuracies. Nonaka et al. [20] further proposed the Brain Hierarchy score (BH score) to quantify such correspondences across models. These works, however, mainly focused on recognition models.

Decoding and reconstruction with generative models. Meanwhile, generative models have been leveraged for decoding and visual reconstruction. Seeliger et al. [13] and Cheng et al. [15] applied generative adversarial networks (GANs) by decoding input vectors from brain activity and feeding them into the model to generate images. Ozcelik and VanRullen [14] decoded stochastic parameters of a variational autoencoder (VAE) and reconstructed images by feeding them into the generative part of the model. While several studies have used generative models for decoding and reconstruction, direct comparison of recognition and generative models in terms of their correspondence with the brain remains limited.

B Details of Methods

This section provides additional methodological details that could not be fully described in the main Methods section.

B.1 fMRI Dataset

We used an fMRI brain activity dataset publicly released by Shen et al. [18] and performed preprocessing following the procedures described in their study. Among the participants in the dataset, we selected three subjects whose performance in the original study was sufficiently high. The dataset consisted of voxel-wise brain activity signals extracted from the participant's visual cortex, with each voxel activity pattern corresponding to either the perception or mental imagery of a single image. For the analysis of each visual area, we used V1, V2, V3, V4, and higher visual cortex (HVC) in ascending order of hierarchy, defined in the same manner as in Shen et al. [11].

The training set comprised brain activity measured while the participant viewed 1,200 natural images. The test set comprised brain activity measured in the perception condition while the participant viewed 50 natural images, and brain activity measured in the imagery condition while the participant imagined 10 natural images and 15 artificial shapes. The natural images in the test set belonged to object categories different from those in the training set.

B.2 Recognition and Generative Models

We adopted the following models as the recognition model and the generative model in this study.

For the recognition model, we used an AlexNet variant known as the BVLC reference CaffeNet [16]. This network consists of five convolutional layers and three fully connected layers, named in order from the input side as conv1, conv2, conv3, conv4, conv5, fc6, fc7, and fc8. In this study, layer numbers corresponded to conv1 = (1), conv2 = (2), conv3 = (3), conv4 = (4), conv5 = (5), fc6 = (6), and fc7 = (7).

For the generative model, we used a generative adversarial network (GAN) provided by Dosovitskiy and Brox [17], which was trained to reconstruct original images from the latent representation at the relu7 layer (i.e., the fc7 layer with a ReLU activation) of the BVLC reference CaffeNet. The network consists of three fully connected layers (defc7, defc6, defc5), three convolutional layers, and five deconvolutional layers, arranged in the following order from input to output: defc7, defc6, defc5, deconv5, conv5, deconv4, conv4, deconv3, conv3, deconv2, deconv1, and deconv0. In this study, the layer numbers for the generative model were assigned as follows: deconv1 = (1), deconv2 = (2), deconv3 = (3), deconv4 = (4), deconv5 = (5), defc6 = (6), and defc7 = (7).

We used seven layers (deconv1–5, defc6, defc7) as targets; intermediate conv layers were not used as decoding targets.

Although this GAN is not perfectly symmetric to the recognition model, it was adopted because of its high fidelity in recovering original images, whereas a fully convolutional autoencoder would yield lower fidelity. While other architectures could be explored in future work, we selected this particular combination of recognition and generative models as a trade-off between architectural symmetry and the generative model's ability to faithfully recover original images. In addition, the layer numbering from (1) to (7) in this study was determined based on the original layer names (e.g., fc6 and defc6 were both assigned as layer (6), and conv3 and deconv3 as layer (3)). Layers sharing the same or similar names were assumed to represent corresponding hierarchical processing stages, and comparisons were made under this assumption.

B.3 Decoder Training and Target Feature Specification

The implementation of the decoders followed the approach of Shen et al. [11]. We trained decoders on the training fMRI dataset to predict the latent representations of recognition and generative models from voxel-wise brain activity patterns. The decoders were implemented as linear prediction using ridge regression, and were trained separately for each of the layers ((1)–(7)) of both models. For each layer l, we estimated the decoder weights $W^{(l)}$ by solving the following ridge regression problem:

$$W^{(l)*} = \arg\min_{W^{(l)}} ||V^{(l)} - UW^{(l)}||_F^2 + \lambda||W^{(l)}||_F^2,$$

where $V^{(l)}$ is the target matrix (number of samples $\times$ number of units in layer l), U is the predictor matrix (number of samples $\times$ number of voxels), $W^{(l)}$ is the weight matrix (number of voxels $\times$ number of units in layer l), and λ is the regularization coefficient ($\lambda = 100$).

The target matrix $V^{(l)}$ is constructed by concatenating the intermediate feature vectors of either recognition model or generative model. The target representations for the recognition model Φ were obtained by inputting an image $\boldsymbol{x}$ into the model and extracting the layer l representation $\Phi_l(\boldsymbol{x})$ where Φ_l represents the function that outputs the intermediate feature vector at the layer index l. In other words, for each row of the target matrix V, the corresponding vector $\boldsymbol{v}$ can be expressed as $\boldsymbol{v} = \Phi_l(\boldsymbol{x})$. For the first layer of the recognition model, we additionally trained decoders using the absolute values of the representations as targets, since we followed the previous studies [10, 20] which evaluated predicted representations with decoders trained on these absolute-valued representations. The target representations for the generative model Ψ were obtained by feeding the relu7 representations of the recognition model as an input vector $\boldsymbol{z} = \mathrm{ReLU}(\Phi_7(\boldsymbol{x}))$ to the generative model Ψ and extracting the corresponding layer l representation $\Psi_l(\boldsymbol{z})$. In other words, for each row of the target matrix $V^{(l)}$, the corresponding vector $\boldsymbol{v}$ can be expressed as $\boldsymbol{v} = \Psi_l(\boldsymbol{z})$. This procedure reflects the relationship between the recognition and generative models described in Appendix B.2.

When training the decoder, we selected the voxels that were considered to contribute most to the prediction of each feature unit, and trained the decoder using only those voxels. For each feature unit, we individually identified the top 500 voxels whose fMRI responses showed the highest correlation with that unit's activity pattern, and used only those voxels to train the decoder.

B.4 Image Reconstruction Procedures

Using the trained decoders, the latent representation at layer l, $\boldsymbol{v}^{(l)}$, was predicted from the brain activity vector $\boldsymbol{u}$ as $\hat{\boldsymbol{v}}^{(l)} = {W^{(l)}}^{\mathsf{T}}\boldsymbol{u}$. Before generating reconstructed images, a scaling was applied to these raw predicted latent representations. For each layer, the standard deviation across units was matched to a specified value. This specified value was the standard deviation computed for each layer from the target representations used while training the decoder. These scaled latent representations are referred to as the predicted latent representations from brain activity, and are used for image reconstruction. In the remainder of this subsection, we use the same symbol $\hat{\boldsymbol{v}}^{(l)}$ for the scaled latent as an abuse of notation.

For the recognition model, reconstructed images were generated using the same procedure as in Shen et al. [11], which applied a Natural Image Prior to reconstruction from a single DNN layer. A randomly initialized input vector was passed through the Natural Image Prior to generate an initial image. The image was then fed into the recognition model to extract the representation of the target layer, and the loss was computed with respect to the representation of that layer predicted from brain activity. The input vector was iteratively updated to minimize this loss, which was based on the squared error:

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} \|\Phi_l(G(\boldsymbol{y})) - \hat{\boldsymbol{v}}^{(l)}\|^2,$$

where $\boldsymbol{y}$ is the input vector of the Natural Image Prior G that outputs an image, $\Phi_l(\boldsymbol{x})$ denotes the vectorized representation of layer l when image $\boldsymbol{x}$ is fed into the recognition model, and $\hat{\boldsymbol{v}}^{(l)}$ is the vectorized predicted

representation of layer l from brain activity. After the optimization, the reconstructed image was obtained as $\hat{\boldsymbol{x}} = G(\boldsymbol{y}^*)$ where $\boldsymbol{y}^*$ is the solution of the above minimization problem. We used the model provided by Dosovitskiy and Brox [17] as the Natural Image Prior, and parameter settings followed the implementation of Shen et al. [11].

For the generative model, the predicted latent representations $\hat{\boldsymbol{v}}$ were directly substituted as the feature representation of the corresponding layer, i.e., $\Psi_l(\boldsymbol{z}) = \hat{\boldsymbol{v}}^{(l)}$, and by forwarding it through the generative model Ψ, the reconstructed image was obtained as $\hat{\boldsymbol{x}} = \Psi(\boldsymbol{z})\Big|_{\Psi_l(\cdot)=\hat{\boldsymbol{v}}^{(l)}}$.

C Additional Results

We present the decoding and reconstruction results using imagery data which are not included in the Results section.

We first decoded hierarchical latent representations (Figure 5a) using imagery data from the whole visual cortex and reconstructed images from the decoded representations. As in the Results section, profile correlation was used to evaluate the decoded representations, and DreamSim similarity was used to evaluate the reconstructed images. The mean profile correlation values for all layers of both models were close to zero, indicating that representations were poorly predicted regardless of the model used (Figure 5b). Although many layers showed higher mean values for the generative model than for the recognition model, these differences were smaller than those observed in the perception condition and did not indicate substantial model-specific effects. Some reconstructed images preserved the shape of the target stimuli for both models, whereas many did not retain any meaningful shape (Figure 5c). While the mean DreamSim similarity values were higher for the generative model than for the recognition model in many layers, overall performance was low, with no notable differences across models or layers.

We then used the imagery data from individual visual areas to decode latent representations and evaluated the hierarchical correspondence of both models with the brain (Figure 5d). As in the Results section, we plotted the proportion of units best predicted by each visual area across model layers and calculated the BH score. Although the BH scores were lower than those for the perception data in both models, there was still evidence of hierarchical correspondence between the models and the brain, especially in the recognition model. Representations from high-level layers were better decoded from higher visual areas for both models. In low-level layers, while representations of the recognition model were better decoded from lower visual areas for the perception data, such regional differences were not observed for the imagery data. For the generative model, as in the perception condition, decoding from low-level layers showed little variation across visual areas.

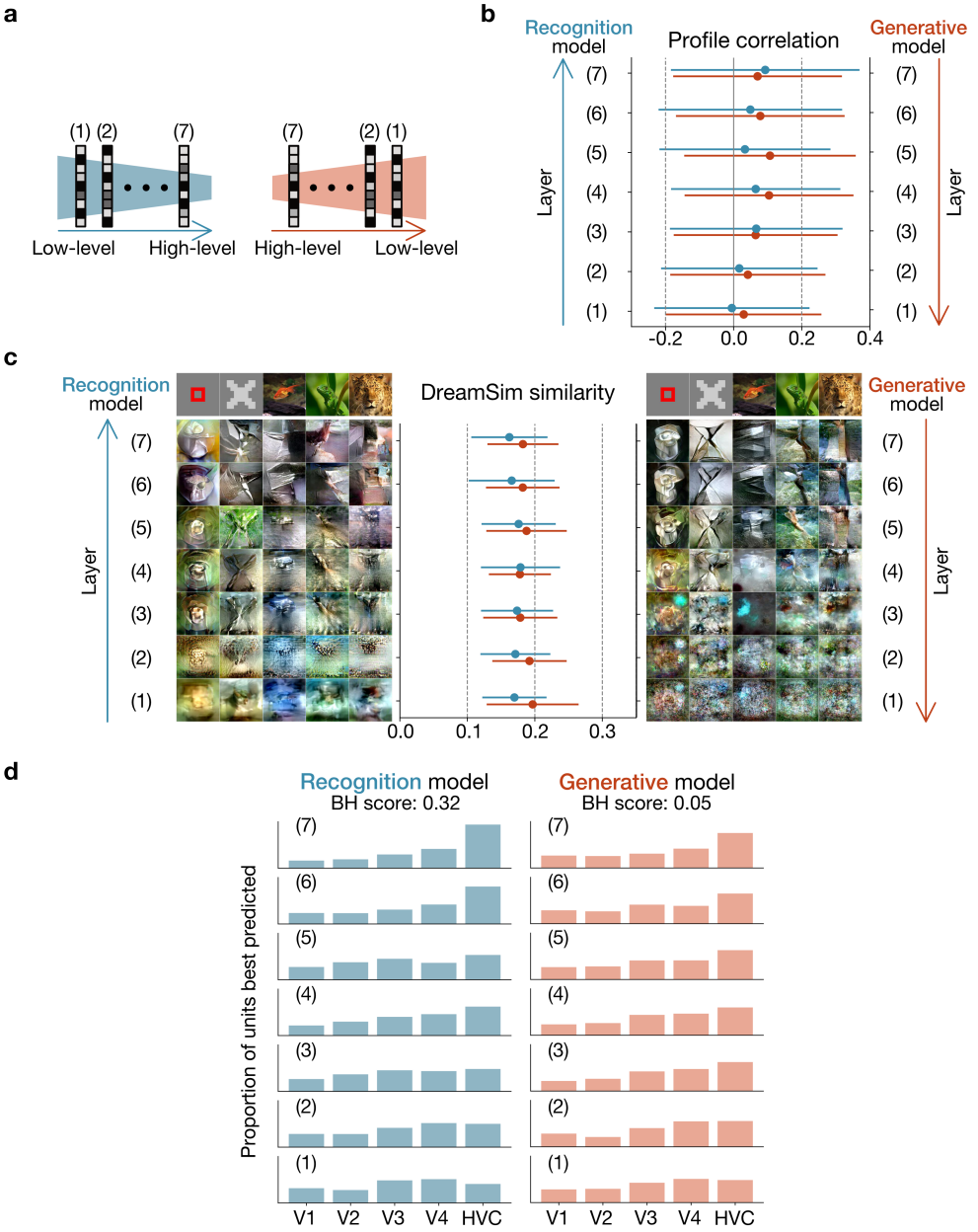

Figure 5: (a) Correspondence between models and layer numbers, increasing from low- to high-level features. (b) Profile correlations of decoded representations from imagery data in the whole visual cortex (mean ± SD across all units pooled from three subjects). (c) Reconstructed images (from a representative subject) and their DreamSim similarity evaluation using the same imagery data (mean ± SD across all stimuli pooled from three subjects). (d) BH scores and the proportion of units best predicted by each visual area across model layers, computed from the imagery data (pooled across three subjects). Visual areas are shown in hierarchical order (V1-V4 and HVC). In all panels, blue denotes the recognition model, and red denotes the generative model.