Decoding Individual Differences in Mental Information from Human Brain Response Predicted by Convolutional Neural Networks

Kiichi Kawahata[1,2], **Jiaxin Wang**[1,2], **Antoine Blanc**[1], **Naoya Maeda**[3],
Shinji Nishimoto[1,2], **Satoshi Nishida**[1,2,*]
[1]Center for Information and Neural Networks (CiNet), Advanced ICT Research Institute,
Natinal Institute of Information and Communications Technology (NICT), Japan
[2]Graduate School of Frontier Biosciences, Osaka University, Japan
[3]NTT DATA Corporation, Japan
[*]s-nishida@nict.go.jp

Abstract

Recent advantages of brain decoding with functional magnetic resonance imaging (fMRI) have enabled us to estimate individual differences in mental information from brain responses to natural sensory inputs. However, the physical constraints and costs of fMRI measurements prevent brain decoding from achieving real-world applications. To address this issue, this study aims to build a framework to decode individual differences in mental information under natural situations via brain-response prediction using convolutional neural networks (CNNs). Once the CNN-based prediction model is constructed using measured brain response, mental information can be decoded from the predicted responses of individual brains with no additional fMRI measurements. As per our analysis, it was found that in 81 of 87 items to be decoded, this framework captured individual difference patterns consistent with conventional decoding using measured brain responses. Our framework has great potential to decode personal mental information with minimal fMRI measuring constraints or costs, which substantially expands the applicability of brain decoding in daily life.

1 Introduction

Brain decoding based on functional magnetic resonance imaging (fMRI) has been identified a valuable tool for not only neuroscience research [7, 26, 27] but also real-world applications, such as neuromarketing [8, 20]. Recent techniques of fMRI-based decoding successfully recovered rich mental information from fMRI signals induced by natural scenes [9, 13, 18, 19, 24]. In addition, several studies have attempted to decode the individual differences in mental information, for example, when personal episodes of daily experiences are recalled [1]. Although these techniques are promising for real-world applications, the physical constraints and the high costs of fMRI measurements prevent these techniques from achieving widespread application.

One approach to solve this problem uses alternative neuroimaging devices, such as electroencephalography (EEG), that require weaker constraint and lower costs [21, 22]. However, signals collected with these devices are noisier than fMRI, thus making it difficult to recover rich mental information [3, 10]. Therefore, the extent of real-world application for brain decoding using these devices remains to be limited. Another approach uses computational methods that can eliminate a large amount of fMRI measurement for brain decoding. An example of such a method is the decoding of fMRI responses predicted by computational models (predicted-response decoding) instead of measured responses [17]. Previous work has demonstrated that this predicted-response decoding achieves the performance comparable to the decoding of measured fMRI responses (measured-response decoding) [17]. Although this method has the potential to drastically reduce the constraints and the costs of fMRI measurements and expand the applicability of brain decoding, the previous work only made the comparison of its group-level performance with measured-response decoding. Thereby its full potential remains unclear.

As an extension of the latter approach, this study aims to introduce predicted-response decoding into the estimation of individual differences in mental information evoked by natural scenes. Inspired by the previously developed method [17], we have constructed prediction models that simulate individual's fMRI responses to arbitrary natural scenes and decoding models that estimate an individual's mental information from the predicted responses (Figure 1). Then, the validity of these models was evaluated using a large variety of decoded items associated with natural scenes, in terms of whether the model estimates accurately captured the individual differences in mental information derived from measured-response decoding. This validation supports the notion that our framework based on predicted-response decoding potentially enables us to achieve weak-constraint, low-cost brain decoding to estimate rich mental information varying from person to person.

2 Predicted-Response Decoding Framework

2.1 Overview

Our framework based on predicted-response decoding consists of models to predict fMRI voxel responses to movie scenes (CNN-to-voxel [cnn2vox] and voxel-to-voxel [vox2vox] models) and to decode movie-associated cognitive labels from the predicted voxel responses (voxel-to-label [vox2lab] models, Figure 1). First, the cnn2vox model transforms the features of movie scenes, extracted via visual and auditory CNNs, to voxel responses. Next, the vox2vox model modifies the predicted voxel response using the history of preceding voxel responses. Finally, the vox2lab model estimates cognitive labels from the modified responses.

The cnn2vox and vox2vox models are pretrained using small datasets of movie-evoked voxel responses collected from individual brains using fMRI. Once the training is completed, these models predict voxel responses to arbitrary movie scenes by transforming the CNN features to the response of individual brains with no additional brain measurement. Then, the vox2lab model is trained using paired datasets of predicted voxel response and cognitive labels linked with movie scenes.

2.2 CNN Feature Extraction

In this study, VGG-16 [25] and SoundNet [2], which are pretrained and available on the web, are used to extract visual and acoustic features, respectively, from movies. To extract visual features from movies via VGG-16 (originally applied to static images with a fixed size of 224 × 224 pixels),

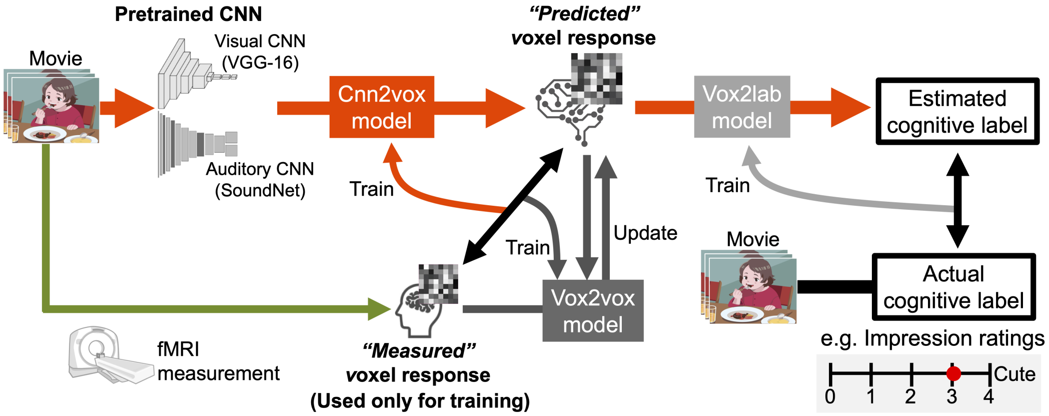

Figure 1: Predicted-response decoding framework.

the movies are decomposed into frames and resized to the same size. Then, unit activations of intermediate layers are calculated when inputting the movie frames and pooled for each second. Finally, the maximum activation value of each unit for each second is used as the visual feature of the movies. This study uses eight layers of pool1–5 and fc6–8 and obtains the visual features for each layer. To extract acoustic features from the same movies, the sound waves of the inputs are resampled with the fixed frequency of 44100 Hz and decomposed into each second. Then, unit activations of intermediate layers in SoundNet are calculated when inputting the sound waves as the acoustic features of the movies. This study uses seven layers of conv1–7 and obtains the acoustic features for each layer. Finally, these processes produce eight series of visual features and seven series of acoustic features of movies.

2.3 Cnn2vox Model

The construction of the cnn2vox, vox2vox, and vox2lab models is based on the voxelwise modeling technique [15]. Using a time series of CNN features and voxel responses, the cnn2vox model acquires the linear mapping from a CNN feature space to a response space of each voxel through statistical learning. The learning objective is to estimate weights of N voxels, as denoted by $\mathbf{W}_{\text{cv}} = \{\mathbf{w}_{\text{cv}(1)}, \cdots, \mathbf{w}_{\text{cv}(N)}\}$, of the linear model: $\mathbf{R} = f(\mathbf{X})\mathbf{W}_{\text{cv}} + \epsilon$, where $\mathbf{R} = \{\mathbf{r}_1, \cdots, \mathbf{r}_N\}$ is a series of responses in each of N voxels, $\mathbf{X}$ is a series of movie scenes, $f(\mathbf{X})$ is its feature representation with the dimensionality of D, and ϵ is isotropic Gaussian noise. A set of linear temporal filters is used to capture the hemodynamic delay in the response [19]. The matrix $f(\mathbf{X})$ is constructed by concatenating four sets of D-dimensional feature vectors with temporal shifts of 3, 4, 5, and 6 seconds. This means that voxel response at a time point t, as denoted by $\mathbf{R}_{(t)}$ $(t = 1, \cdots, T)$, is modeled by a weighted linear combination of the preceded series of CNN features:

$$\mathbf{R}_{(t)} = \sum_{k=3,4,5,6} f(\mathbf{X}_{(t-k)})\mathbf{W}_{\text{cv},k} + \epsilon,$$

where $\mathbf{W}_{\text{cv},k}$ denotes the weights corresponding to the delay k. The weight estimation is performed using L2-regularized linear least-squares regression. The optimal regularization parameter for each model is determined by 10-fold cross validation of training data and shared across all voxels. In this study, $f(\mathbf{X})$ represents a series of unit activations induced by movies for each layer of VGG-16 or SoundNet. Since the substantial number of units in lower layers of the CNNs took too much computational cost for the regression process, the dimensionality of unit-activation features for each layer is reduced in advance by principal component analysis on training datasets. This study reduces the dimensionality, D, to 1000. Finally, eight models for eight VGG-16 layers (pool1–5 and fc6–8) and seven models for seven SoundNet layers (conv1–7) are constructed for each brain.

Each of the estimated linear models predicts voxel responses to new movie scenes. Then, the predicted voxel responses from individual models are integrated using linearly weighted averages for each voxel. The weight for a given voxel is calculated based on the prediction accuracy (Pearson correlation coefficient between measured and predicted voxel responses) for that voxel calculated during the cross-validation in model training. Specifically, the weight for the i-th model, as denoted by w_i, is determined by $w_i = a_i / \sum_j^{15} a_j$, where a_i is the prediction accuracy of the model. This integration process produces a single series of predicted voxel responses to the new movie scenes.

2.4 Vox2vox Model

The vox2vox model predicts response in one voxel at a given time point from responses in a group of voxels at preceding time points. Hence, this model captures endogenous properties of voxel responses, such as intrinsic connectivity between different brain regions [4]. Meanwhile, the cnn2vox model captures exogenous properties of voxel responses, such as stimulus selectivity. The vox2vox model was reported to improve performance in predicting voxel response to movie scenes [17].

In the vox2vox model, a response in each of N voxels at a time point t, as denoted by $\mathbf{R}_{(t)}$, is modeled by a weighted linear combination of responses in the set of M voxels preceded by 1, 2, and 3 seconds:

$$\mathbf{R}_{(t)} = \sum_{k=1,2,3} \mathbf{R}'_{(t-k)}\mathbf{W}_{\text{vv},k} + \epsilon.$$

The M voxels are selected based on the cnn2vox model prediction accuracy on model training data. In this study, the top 2000 voxels with the highest prediction accuracies after the weighted average of all the models are used as the M voxels. The regression procedure is the same as that used in the cnn2vox model. Response predictions from the cnn2vox and vox2vox models are thereafter combined as a weighted sum. The weight is determined by the relative accuracy of each prediction for each voxel.

2.5 Vox2lab Model

The vox2lab model estimates cognitive labels associated with movie scenes from predicted voxel responses. In this model, a series of z-scored cognitive labels at a time point t, denoted by $\mathbf{L}_{(t)}$ $(t = 1, \cdots, T')$, is regressed by a series of predicted responses to the scenes in the set of N voxels with the hemodynamic delay, k, of 3, 4, and 5 seconds [18]:

$$\mathbf{L}_{(t)} = \sum_{k=3,4,5} \hat{\mathbf{R}}_{(t+k)} \mathbf{W}_{\mathrm{vl},k} + \epsilon.$$

The regression procedure is the same as that used in the other models, except that the regularization parameters are determined separately for each dimension of label vectors. This model learns the association between predicted voxel responses (but not measured voxel responses) and cognitive labels.

3 Data

3.1 Movie

Two sets of movies were used for experiments and analysis. One includes 368 Japanese ad movies broadcasted on the web between 2015 and 2018 (web ad movies). The other includes 420 Japanese ad movies broadcasted on TV between 2011 and 2017 (TV ad movies). The movies are all unique, include a wide variety of product categories (Table S1), and have the same resolution (1280 $\times$ 720 pixels) and frame rate (30 Hz). The lengths of the movies are either 15 or 30 seconds. They are also accompanied by PCM sounds with the sampling rate of 44100 Hz and are normalized so that they have the same RMS level. To create movie stimuli for experiments, the original movies in each set were sequentially concatenated in a pseudo-random order. Each stimulus set of the movies has the length of 8400 seconds in total and was divided into 7200 seconds and 1200 seconds to collect voxel responses for the training (training dataset) and test (test dataset), respectively, of all the three models.

3.2 fMRI Data

fMRI responses to the movie stimuli were collected from Japanese participants using a 3T MRI scanner with a sampling rate of 1 Hz. In total, 40 (15 females; age [mean ± SD] = 26.6 ± 9.0 years) and 28 (12 females; age [mean ± SD] = 26.4 ± 7.6) participants were assigned to the fMRI experiments with the web ad movie set and those with the TV ad movie set, respectively. Of these, 16 participants overlapped between the two experiments. The experimental protocol is approved by the ethics and safety committees of the authors' institution. Written informed consent was obtained from all participants. For the modeling in each participant, the fMRI data were preprocessed and all voxels within the whole cortex were extracted. For more details, see Supplementary Methods.

3.3 Cognitive Labels

The five categories of cognitive labels associated with movie scenes used to assess the validity of predicted-response decoding are as follows: (1) scene descriptions, (2) impression ratings, (3) ad effectiveness indices, (4) ad preference votes, and (5) subjective preference ratings. The categories (1) and (2) are linked to both movie sets. The category (3) is linked to the web ad movie set. The categories (4) and (5) are linked to the TV ad movie set. The category (5) was collected fully from the fMRI participants. Each of the categories (2)–(4) has subordinate items. In total, 87 labels are assigned to the time series of the movie sets. The details of each category are described below. For additional details, see Supplementary Methods and Table S2.

Scene Descriptions Manual descriptions given for every 1-second movie scene were collected from human annotators who were instructed to describe each scene. The descriptions contain a variety of expressions reflecting not only their objective perceptions but also their subjective perceptions (e.g., feeling). To evaluate the scene descriptions quantitatively, the descriptions were transformed into vectors of word2vec [14]. Individual words in each description were projected into the pretrained word2vec vector space. Then, the word vectors obtained from all descriptions within each scene were averaged. This procedure yielded one 100-dimensional vector for each 1-second scene. Thus, one label set was assigned to each movie set.

Impression Ratings Manual ratings given for every 2-second movie scene on 30 different impression items (e.g., ”beautiful”) were collected from human annotators. While the annotators sequentially watched separate 2-second clips of the movies, they evaluated each item on a 5-point scale from 0 to 4. The mean impression ratings in every 2-second scene were obtained by averaging multiple ratings and then oversampled to obtain time series of rating labels in every 1-second scene. Thus, 30 label sets were assigned to each movie set.

Ad Effectiveness Indices Two types of mass behavior indices were collected from the Internet. One type is click-through rate, or the fraction of viewers who clicked the frame of a movie and jumped to a linked web page. The other type is view completion rate, or the fraction of viewers who continued to watch an ad movie until a specific time point of the movie (25%, 50%, 75%, or 100% from the start) without choosing a skip option. Hence, each movie has four indices of the view completion rate. Although a single value of each index was assigned to each ad, time series of indices in every 1-second scene were obtained by filling all scenes in an ad with an identical index value assigned to the ad. Thus, five label sets were assigned to the web ad movie set.

Ad Preference Votes Reputation surveys of TV ads for commercial purposes were conducted using questionnaires to large-scale testers. Each tester was asked to freely recall a small number of her/his favorite TV ads from among the ads recently broadcasted. The total number of recalls of an ad was regarded as the preference vote. In addition, the questionnaires also include 15 subordinate items that ask why the ads are favorable for her/him (e.g., ”humorous”). Three additional items reflected how effective the ads were for her/his usage and purchase of products (e.g., ”purchase intention”). Although one value of each item was assigned to each ad, the time series of the value in every 1-second scene were obtained by filling all scenes in an ad with an identical value assigned to the ad. Since these values are distributed in a similar form of a gamma distribution, the logarithm of the data was taken. Thus, 19 label sets were assigned to the TV ad movie set.

Subjective Preference Ratings Manual preference ratings given for each TV ad movie were collected from 14 of the fMRI participants. While the participants sequentially watched each movie outside the MRI scanner, they rated their own preference for the movie on a 9-point scale from -4 to 4. The ratings were oversampled to obtain a time series of ratings in every 1-second scene. Thus, one label set was assigned to the TV ad movie set.

4 Analysis

4.1 Model Construction

For the predicted-response decoding, the three models for each participant are constructed using the pair data of web or TV ad movies, the participant's fMRI data, and cognitive labels in the training dataset. After the cnn2vox and vox2vox models are trained using movies and the measured voxel responses to them, the vox2lab model for each cognitive label is trained using the predicted voxel responses to the same movies and the label associated with them. Then, a time series of each label is estimated using these three models with the test dataset.

The Pearson correlation coefficient between this estimated time series and a time series of the true cognitive label is calculated for each participant as the measure of decoding accuracy for each label. Note that since the label of scene description uses multidimensional vectors, the decoding accuracy for this label is evaluated by calculating the correlation coefficient at each time point and averaging it over all time points. Whether the decoding accuracy for a given label is significantly higher than zero is evaluated with the Wilcoxon signed-rank test with correction for multiple comparisons using the false discovery rate (FDR), drawing on each participant's accuracy as a data sample ($P < 0.05$).

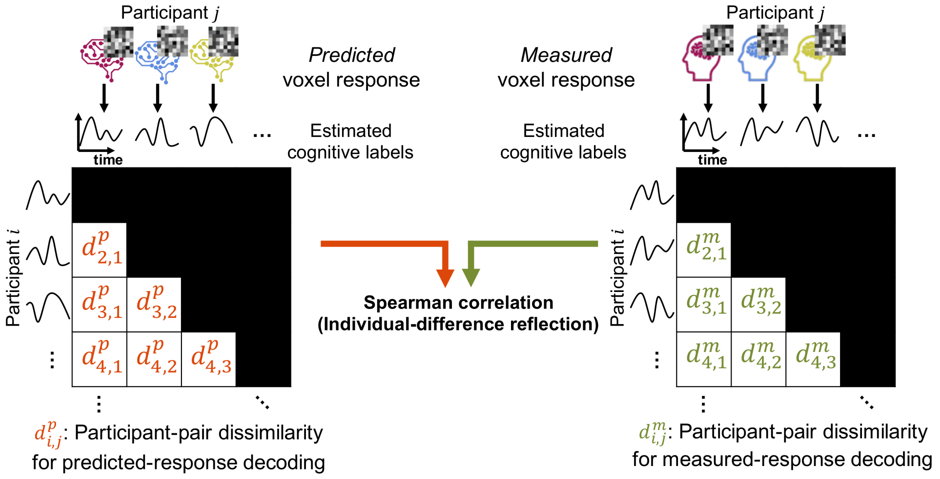

Figure 2: Evaluation of individual-difference reflection.

For the comparison with the predicted-response decoding, this study has also constructed each participant's model for the measured-response decoding, in which cognitive labels are directly estimated by decoding measured voxel responses to movies. The form of this model is the same as in the vox2lab model of the predicted-response decoding, except that measured voxel responses are used instead of predicted responses. The regression procedure and the significance test are the same as those used in the predicted-response decoding.

4.2 Individual Difference

The individual differences of estimation from each of the two decoding methods is evaluated using the participant-pair dissimilarity of the time series of decoded cognitive labels for the test dataset (Figure 2). The dissimilarity is measured with the Pearson correlation distance (i.e., 1 − Pearson correlation coefficient) of the time series between all possible pairs of the participants for each movie set. Then, the Spearman correlation of participant-pair dissimilarities between predicted and measured-response decoding is calculated for each label as a measure of how well the individual differences derived from the predicted-response decoding reflect those derived from the measured-response decoding (referred to as *individual-difference reflection*). Note, that this study uses Spearman correlation as previously recommended for computing the correlation of dissimilarity measure [16]. However, comparable results are observed even when Pearson correlation was used (see Section 5.2). When the Spearman correlation for a given label is significantly larger than 0 ($P < 0.05$, FDR corrected), the predicted-response decoding is regarded as reflective of the individual differences in mental information for that label.

5 Results

5.1 Model Performance

To validate the models constructed for the predicted-response decoding, we first examine the model performance in terms of voxel-response prediction and decoding at the population level. The accuracy of voxel-response prediction (prediction accuracy) in the cnn2vox models is evaluated by the Pearson correlation between predicted and measured voxel responses. The localized pattern for the prediction accuracy of the cnn2vox model across the cortex shows that the models based on the visual and auditory CNNs selectively predict voxel responses in visual and auditory cortical regions, respectively (Figure S1). This is consistent with previous findings in brain-response modeling studies using CNNs [5, 6, 11].

The decoding accuracy of the vox2lab model (see Section 4.1) is significantly higher than 0 for all 87 cognitive labels (0.11–0.82; Wilcoxon signed-rank test, $P < 0.0005$, FDR corrected) and strongly correlated with that of measured-response decoding across all labels (Pearson $r = 0.69$, $P < 0.0001$;

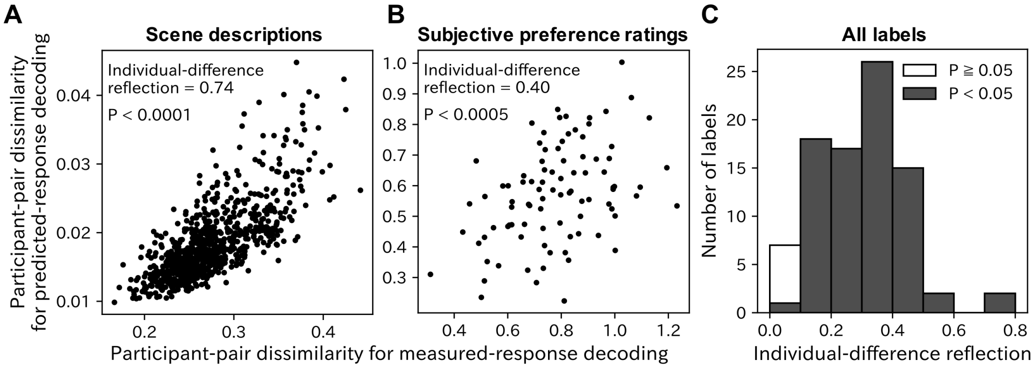

Figure 3: Individual-difference reflection for two example labels (A and B) and for all 87 labels (C).

Figure S2). Furthermore, the accuracy is rather higher for the predicted-response decoding than for measured-response decoding (Wilcoxon signed-rank test, P < 0.0001) as reported previously [17]. Thus, the models effectively work for estimating mental information, at least, at the population level.

5.2 Significant Individual-Difference Reflection for Almost All Cognitive Labels

We next examine how well the predicted-response decoding captures the individual differences in mental information derived from the measured-response decoding. Among all cognitive labels, the highest value of the individual-difference reflection (see Section 4.2) is observed in the label of scene descriptions for the web ad movies (0.74; Figure 3A). The estimation of this label is derived from the pair data of predicted voxel response and cognitive label each of which is derived from different persons. However, even when using the pair data derived from the same person (i.e., when estimating the label of subjective preference ratings), we obtain a significant value of the individual-difference reflection (P < 0.0005; Figure 3B). In addition, the individual differences of this label estimated from the predicted-response decoding are also significantly correlated with the individual differences of preference ratings themselves (Figure S3). These results suggest that the predicted-response decoding successfully captures the individual differences in mental information associated with behavior.

Of all the 87 cognitive labels to be decoded, we observe significant individual-difference reflection for 81 labels (Figure 3C and Table S3; P < 0.05, FDR corrected). A similar result is observed even when a different measure of individual-difference reflection is used (i.e., Pearson correlation instead of Spearman correlation; Figure S4). Thus, these results suggest that the predicted-response decoding successfully captures the individual differences in mental information for almost all the cognitive labels used across different datasets.

5.3 Relationship Between Individual-Difference Reflection and Decoding Performance

The individual-difference reflection, however, varies across cognitive labels. This raises the question of what factor determines the variation of the individual-difference reflection. To address this, we hypothesize that the degree of the individual-difference reflection for a given label is affected by the accuracy of the predicted-response decoding for that label. A significant correlation between the individual-difference reflection and decoding accuracy indicates the validity of this hypothesis (Pearson r = 0.49, P < 0.0005; Figure S5). Therefore, the more accurately we decode a cognitive label using the predicted-response decoding, the more effectively the decoded contents reflect the individual differences in mental information for the label.

6 Conclusion

In this study, we have aimed to build the framework of predicted-response decoding to estimate the individual differences in mental information evoked by natural scenes with minimal fMRI measurement. As per our findings, it was determined that this framework was able to successfully captures the individual-difference patterns of mental information, derived from conventional measured-response decoding, for 81/87 (93.1%) of decoded cognitive labels associated with natural scenes. Our findings suggest that this framework can be used to estimate personal mental information evoked by natural scenes.

We examined the predicted-response decoding framework using moderate-performance CNNs (i.e., VGG-16 and SoundNet) for extracting features from movies and using linear regression for mapping brain responses with movie features or cognitive labels, according to the settings in a previous study [17]. Although the framework successfully captures the individual differences in mental information even for this setting, the reflection of the individual differences may be improved by employing higher-performance CNNs [23] and/or more sophisticated nonlinear regression (e.g., regression with deep neural networks [12]). It is important to note that our framework essentially allows us to apply any methods to feature extraction and regression. Therefore, the optimal component methods for the predicted-response decoding of individual differences should be addressed in future work.

We have confirmed that predicted-response decoding can capture the individual differences of not only mental information, derived from measured-response decoding, but also the behavior of subjective preference ratings for movies (Figure S3). This result suggests that the individual differences estimated by predicted-response decoding potentially reflect subjective perception and cognition of individuals during natural stimulation. This is similar to sensory input given in everyday situations. Therefore, predicted-response decoding provides a low-cost, versatile tool for brain decoding to estimate personal perception and cognition evoked by natural scenes, which has immense potential to facilitate real-world applications of brain decoding.

However, the potential negative societal impacts of predicted-response decoding may include involuntary mind reading and negative bias against personality. Predicted-response decoding can evaluate individual's mental information evoked by natural scenes with no additional brain measurement. Therefore, someone's mental information may be evaluated by a third party without her/his realizing. In addition, since specific types of the evaluated mental information (e.g., subjective preference) are related to her/his personality, the evaluation may in some cases prejudice her/his personality. Although predicted-response decoding may not yet have achieved the level of performance that raises such concerns, the future development of this method will also need efforts to address how to ensure the security of personal models and how to define the correct use of personal models by law.

References

[1] Andrew James Anderson, Kelsey McDermott, Brian Rooks, Kathi L. Heffner, David Dodell-Feder, and Feng V. Lin. Decoding individual identity from brain activity elicited in imagining common experiences. *Nature Communications*, 11(1):1–14, 2020.

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016.

[3] Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A. Robbins. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, 9:16, Jun 2015.

[4] Michael D Fox and Marcus E Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews. Neuroscience*, 8(9):700–711, Sep 2007.

[5] Umut Güçlü and Marcel A. J. van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014, Jul 2015.

[6] Umut Güçlü and Marcel A.J. van Gerven. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336, Jan 2017.

[7] James V Haxby, Andrew C Connolly, and J Swaroop Guntupalli. Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, 37:435–456, Jun 2014.

[8] Lin He, Thomas Freudenreich, Wenhuan Yu, Matthew Pelowski, and Tao Liu. Methodological structure for future consumer neuroscience research. *Psychology Marketing*, 38(8):1161–1181, Aug 2021.

[9] Alexander G. Huth, Tyler Lee, Shinji Nishimoto, Natalia Y. Bilenko, An T. Vu, and Jack L. Gallant. Decoding the semantic content of natural movies from human brain activity. *Frontiers in Systems Neuroscience*, 10(October):1–16, 2016.

[10] Mainak Jas, Denis A. Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159:417–429, Oct 2017.

[11] Alexander J.E. Kell, Daniel L.K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.e16, May 2018.

[12] Meenakshi Khosla, Gia H Ngo, Keith Jamison, Amy Kuceyeski, and Mert R Sabuncu. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances*, 7(22):2020.09.11.293878, May 2021.

[13] Eri Matsuo, Ichiro Kobayashi, Shinji Nishimoto, Satoshi Nishida, and Hideki Asoh. Generating natural language descriptions for semantic representations of human brain activity. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 22–29, 2016.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119, 2013.

[15] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, May 2011.

[16] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553, 2014.

[17] Satoshi Nishida, Yusuke Nakano, Antoine Blanc, Naoya Maeda, Masataka Kado, and Shinji Nishimoto. Brain-mediated transfer learning of convolutional neural networks. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, pages 5281–5288, 2020.

[18] Satoshi Nishida and Shinji Nishimoto. Decoding naturalistic experiences from human brain activity via distributed representations of words. *NeuroImage*, 180(A):232–242, 2018.

[19] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, Oct 2011.

[20] Hilke Plassmann, Vinod Venkatraman, Scott Huettel, and Carolyn Yoon. Consumer Neuroscience: Applications, Challenges, and Possible Solutions. *Journal of Marketing Research*, 52(4):427–435, Aug 2015.

[21] Mamunur Rashid, Norizam Sulaiman, Anwar P. P. Abdul Majeed, Rabiu Muazu Musa, Ahmad Fakhri Ab. Nasir, Bifta Sama Bari, and Sabira Khatun. Current Status, Challenges, and Possible Solutions of EEG-Based Brain-Computer Interface: A Comprehensive Review. *Frontiers in Neurorobotics*, 14:25, Jun 2020.

[22] Maham Saeidi, Waldemar Karwowski, Farzad V. Farahani, Krzysztof Fiok, Redha Taiar, P. A. Hancock, and Awad Al-Juaid. Neural Decoding of EEG Signals with Machine Learning: A Systematic Review. *Brain Sciences*, 11(11):1525, Nov 2021.

[23] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*, 2018.

[24] K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, and M.A.J. van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, Nov 2018.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, page 1409.1556, Sep 2014.

[26] Frank Tong and Michael S Pratte. Decoding patterns of human brain activity. *Annual review of psychology*, 63:483–509, 2012.

[27] Marcel A. J. van Gerven, Katja Seeliger, Umut Güçlü, and Yağmur Güçlütürk. Current Advances in Neural Decoding. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 379–394. Springer, 2019.

S1 Supplementary Methods

S1.1 fMRI Data Collection and Preprocessing

A 3T Siemens MAGNETOM Prisma scanner was used with a 32-channel Siemens volume coil and a multiband gradient echo-EPI sequence [5] (TR = 1000 ms; TE = 30 ms; flip angle = 60°; voxel size = $2 \times 2 \times 2$ mm; matrix size = 96×96; the number of axial slices = 72; multiband factor = 6). The field of view covered the entire cortex. Using a T1-weighted MPRAGE sequence on the same 3T scanner, anatomical data were also collected (TR = 2530 ms; TE = 3.26ms; flip angle = 9° ; voxel size = $1 \times 1 \times 1$ mm; matrix size = 256×256; the number of axial slices = 208).

In these experiments, the participants viewed the movie stimuli on a projector screen inside the scanner ($27.9° \times 15.5°$ of visual angle) and listened to the accompanying sounds through MR-compatible headphones. The participants were given no explicit task. The fMRI response data for each set of the web and TV ad movies were collected from individual participants in three separate recording sessions over 3 days.

For each stimulus set, 14 non-overlapping movie clips of 610 seconds in length were obtained. The individual movie clips were displayed in separate scans. The initial 10-second part of each clip was a dummy to discard hemodynamic transients caused by clip onset. The fMRI responses collected during the 10-second dummy movie were not used for the modeling. The 12 clips, including movies of the training dataset, were presented once. The fMRI responses to these clips are used for model training (7200 seconds in total). The other clips including movies of the test dataset were presented four times each in four separate scans. The fMRI responses to these clips were averaged across four scans and used for model test (1200 seconds in total).

For fMRI data preprocessing, motion correction in each functional scan was performed using the statistical parameter mapping toolbox (SPM8, `http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`). For each participant, all volumes were aligned to the first image from the first functional run. Using a median filter with a 120-second window and subtracting from the signal, low-frequency fMRI response drift was detected. Then, the response for each voxel was normalized by subtracting the mean response and scaling it to the unit variance. FreeSurfer [1, 2] was used to identify cortical surfaces from anatomical data and register them to the voxels of functional data. All voxels identified within the whole cortex for each participant were used for the modeling. In addition, cortical voxels were anatomically segmented into 358 regions based on the HCP-MMP1 atlas [3] in order to show the localization of model performance on predicting voxel response.

S1.2 Details of the Cognitive Labels

In this study, five distinct categories of cognitive labels associated with movie scenes are used to evaluate the validity of the predicted-response decoding. The details of each category are described in the following and summarized in Table S2.

1. Scene descriptions Manual scene descriptions using natural Japanese language were provided for every 1-second scene of each of the web and TV ad movie sets. The annotators were native Japanese speakers (web ad movies: 11 females and 2 males, age 20–56 years; TV ad movies: 68 females and 28 males, age 19–62 years), who were not the fMRI participants. They were instructed to describe each scene (the middle frame of each 1-second clip) using more than 50 Japanese characters. Multiple annotators (web ad movies, 5 annotators; TV ad movies: 12–14 annotators) were randomly assigned for each scene to reduce the potential effect of personal bias. The descriptions contain a variety of expressions reflecting not only objective perceptions but also the subjective perceptions of the annotator (e.g., impression, feeling, association with ideas). The following are the descriptions of two example scenes:

Example scene 1

- In the middle of a light blue background, there is a man who has a tennis racket in his hand and glares at the front. The scene tells me he is very serious.
- He is a tennis player who everyone really admires. He has a cool face. He is handsome and refreshing. It looks like an advertisement for men.

- A short-haired man with a suntanned skin looks at a point with a sharp look. He has a racket in his hand. He is wearing blue clothes.
- The power of his eyes is amazing. His sharp eyes give me the impression that he intimidates his opponent. As a sportsman, he seems to be clean even if he is sweating.
- A young Asian man with short black hair faces his eyes forward. I can see his racket. There are white letters in the lower part of the screen.

Example scene 2

- It looks like a Christmas scene. It seems like Santa Claus is going back to the sky. It looks good for a Christmas gift.
- A book on the green cover is opened. The book is on the brown table. In the book, there is a round green character waving its hands.
- A thick open book with a green spine cover is placed on the table of brown trees. It looks like a very funny picture book.
- A green book is opened on the desk of the tree. On the opened page, a moon is drawn with a round green stuffed animal waving toward the moon.
- It is cute. It seems to be waving bye-bye. Before Christmas? It seems like he is looking down on Mr. Reindeer and Santa.

To evaluate the scene descriptions quantitatively, the descriptions were transformed into vectors of word2vec [4] in the same way as in a previous work [6]. The word2vec (skip-gram) algorithm was originally developed to learn a high-dimensional word vector space based on local (nearby) word co-occurrence statistics in natural language texts. A word2vec vector space was constructed ahead from a text corpus of the Japanese Wikipedia dump on January 11, 2016 (vector dimensionality = 100; window size = 10; the number of negative samples = 5). All Japanese texts in the corpus were segmented into words and lemmatized using MeCab (`http://taku910.github.io/mecab`). The only parts of speech used were nouns, verbs, and adjectives. All the others were discarded. To improve the reliability of the word2vec learning and restrict the vocabulary size to around 100000 words, words that appeared less than 178 times in the corpus were excluded.

Each description for a given scene was also segmented, lemmatized, and decomposed into nouns, verbs, and adjectives using MeCab as has been described above. Individual words were projected into the word2vec vector space. The word vectors were averaged within each description. Then, for each scene, all vectors that were obtained from the different descriptions were averaged. This procedure yielded one 100-dimensional vector (description vector) for each 1-second scene.

2. Impression ratings Manual ratings of 30 different impressions were provided for every 2-second scene of each of the web and TV ad movie sets. The annotators for the web ad movie sets were completely different from the fMRI participants (26 females and 6 males: age 24–61). However, the annotators for the TV ad movie set were partly recruited from the fMRI participants (3 females and 3 males from the fMRI participants, age 22–46 years: 5 females and 1 male from others, age 20–45 years). The ratings were given for every 2-second scene of movies. All the impression items are shown in Table S2. While the annotators sequentially watched separate 2-second clips of the movies with sounds, they evaluated each item on a 5-point scale from 0 to 4. To keep the motivation of the annotators, only 15 labels were assigned to each annotator in single clip evaluation. All the items were evaluated from 9 different annotators for each scene of the web ad movies and from 12 different annotators for each scene of the TV ad movies. The mean impression ratings in every 2-second scene were obtained by averaging all the ratings for each movie and then oversampled to obtain a time series of rating labels in every 1-second scene.

3. Ad effectiveness indices To evaluate the effectiveness of individual web ad movies, two types of mass behavior indices were collected on the web. One type is click-through rate, defined as the fraction of viewers who clicked the frame of a movie and jumped to a linked web page. The other type is view completion rate, or the fraction of viewers who continued to watch an ad movie until a specific time point of the movie (25%, 50%, 75%, or 100%) without choosing a skip option (hence, there are four indices for each movie). These 5 indices were computed from 3783 to 12268028 (mean = 632516) unique accesses for each movie. Although a single value of each index was assigned to each movie clip, time series of indices in every 1-second scene were obtained by filling all 1-second scenes in a clip with an identical index value assigned to the clip.

4. Ad preference votes Mass preference votes for each clip in the TV ad movies were collected for commercial investigation using questionnaires conducted on large-scale testers. In the questionnaires, each tester was asked to freely recall a small number of her/his favorite TV ads for her/him from among the ads recently broadcasted on TV. In addition, the questionnaires also include 15 subordinate items that asked why the ads were favorable for her/him (e.g., ”humorous”). Three additional items reflected how effective the ads were for her/his usage and purchase of products (e.g., ”purchase intention”: Table S2). To eliminate the bias of preference due to frequent exposure on TV, the number of votes for each item of an ad was divided by its gross rating point (GRP). GRP is an index of how many people see an ad over a particular period. It is calculated by the audience ratio multiplied by the number of times the ad is broadcasted during the period. Although a single value for each item was assigned to each movie clip, a time series of the value in every 1-second scene was obtained by filling all 1-second scenes in a clip with an identical value assigned to the clip.

5. Subjective preference ratings Manual preference ratings given for each TV ad movie were collected from 14 of the fMRI participants (9 females and 5 males: age 21–49) on separate days after their fMRI experiments. While the participants sequentially watched each movie presented on a computer screen outside the MRI scanner, they rated their own preference for the movie on a 9-point scale from -4 (most unlikable) to 4 (most likable). To keep the motivation of the participants, the ratings were conducted in separate seven blocks (600 seconds per block) with intervals on each of two days. The values collected in the ratings were oversampled to obtain a time series of ratings in every 1-second scene.

Supplementary References

[1] Anders M. Dale, Bruce Fischl, and Martin I Sereno. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, Feb 1999.

[2] Bruce Fischl, Martin I Sereno, and Anders M Dale. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, Feb 1999.

[3] Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, Aug 2016.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119, 2013.

[5] Steen Moeller, Essa Yacoub, Cheryl A. Olman, Edward Auerbach, John Strupp, Noam Harel, and Kâmil Uğurbil. Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic Resonance in Medicine*, 63(5):1144–1153, May 2010.

[6] Satoshi Nishida and Shinji Nishimoto. Decoding naturalistic experiences from human brain activity via distributed representations of words. *NeuroImage*, 180(A):232–242, 2018.

S2 Supplementary figures

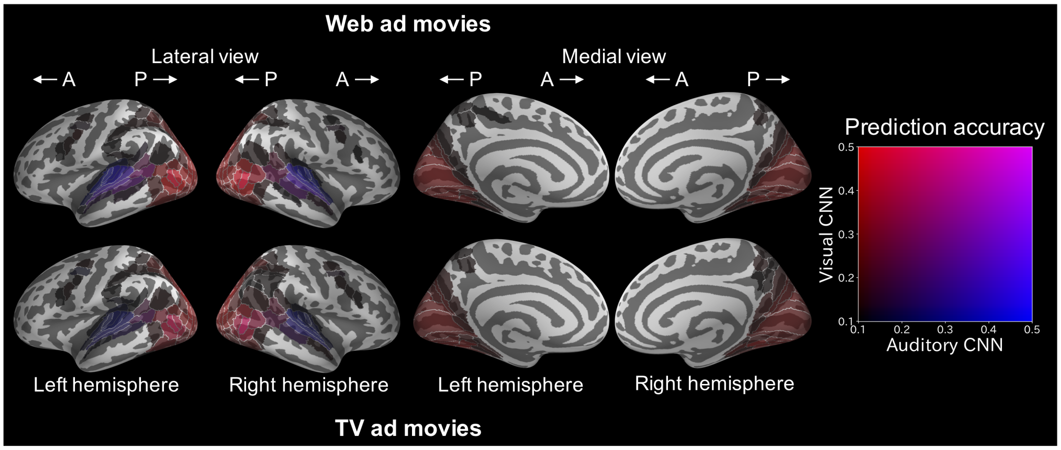

Figure S1: Localization of cortical regions in which the response is predicted by cnn2vox models. Participant-average prediction accuracy is shown within each of the 358 cortical regions segmented with the HCP-MMP1 atlas. Colors indicate the mean prediction accuracy within each of the regions denoted by the colormap (right). Accuracy less than 0.1 is not shown. The regions predicted preferentially by the visual CNN-based models and the regions predicted preferentially by the auditory CNN-based models are shown in distinct colors (red and blue, respectively). The visual CNN-based models accurately predict voxel responses in visual regions, such as occipital and posterior temporal cortical areas. In contrast, the auditory CNN-based models predict voxel responses in auditory regions, such as anterior temporal cortical areas; A, anterior; P, posterior.

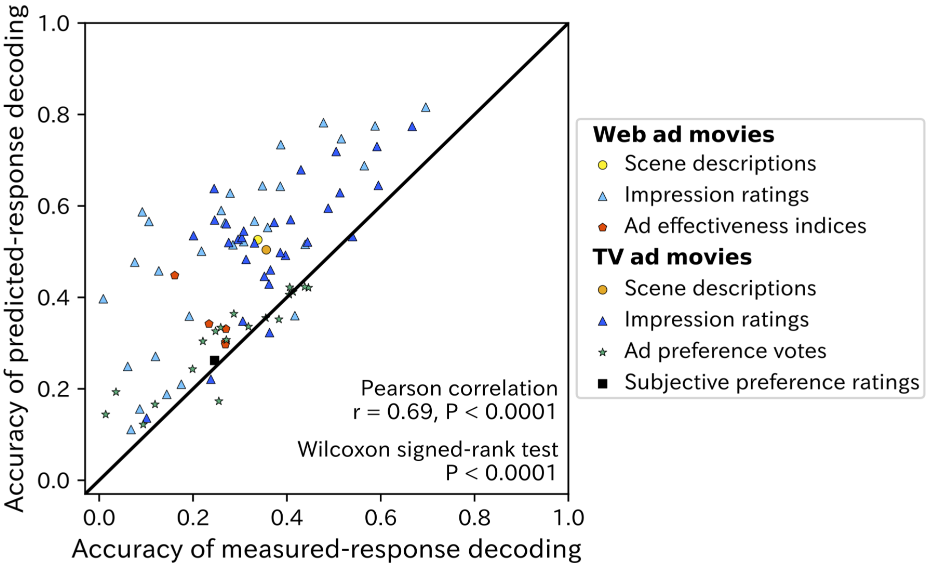

Figure S2: Accuracy of predicted- and measured-response decoding. Each dot denotes accuracy for each cognitive label. Blue and red dots are derived from web and TV ad movie sets, respectively. The accuracy is significantly correlated between these two methods (Pearson r = 0.69; P < 0.0001). The accuracy of predicted-response decoding is significantly higher than that of measured-response decoding (Wilcoxon signed-rank test, P < 0.0001).

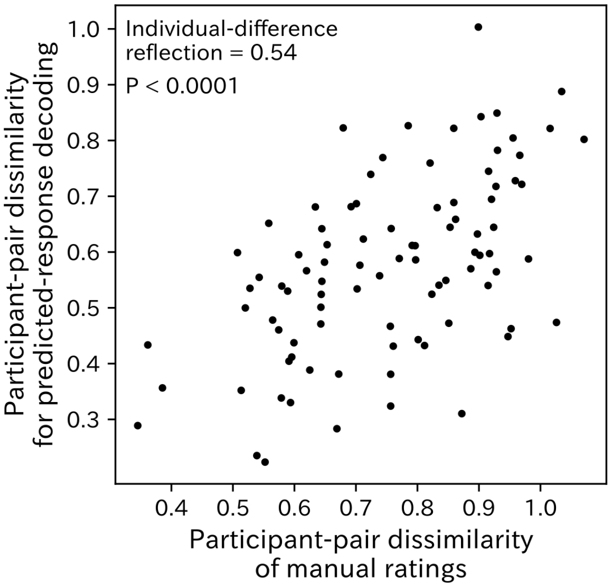

Figure S3: Correlation between the individual differences of subjective preference ratings estimated from predicted-response decoding and those of manual ratings themselves. Each dot denotes the dissimilarity of each participant pair. The correlation (individual-difference reflection) is significant (0.54; P < 0.0001), suggesting that the predicted-response decoding can capture the individual differences even in mental information associated with behavior.

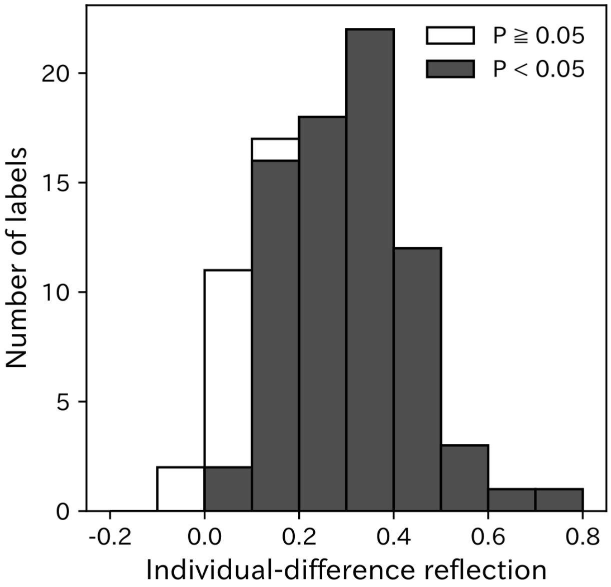

Figure S4: Individual-difference reflection for all cognitive labels when using Pearson correlation instead of Spearman correlation. In this case, 75 of 87 labels show significant individual-difference reflection (gray bars; P < 0.05, FDR corrected).

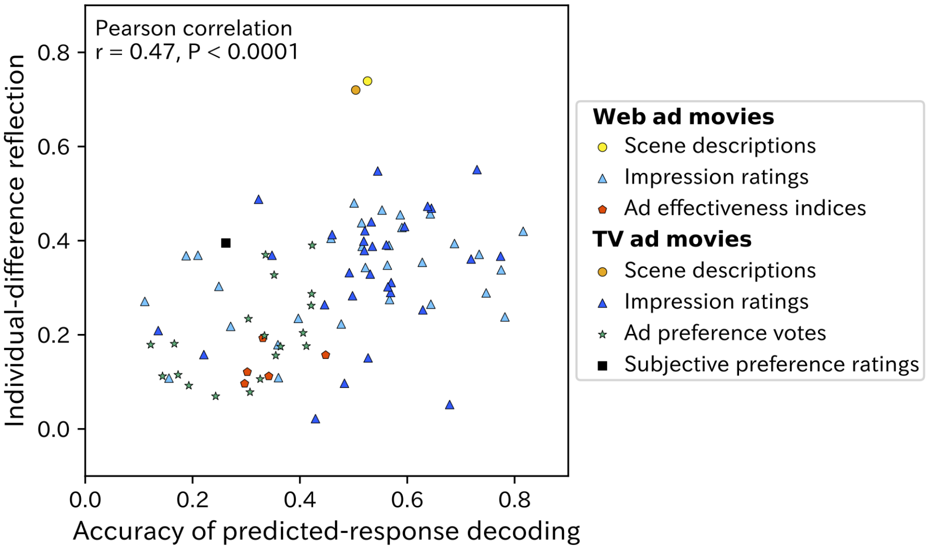

Figure S5: Correlation between individual-difference reflection and the accuracy of predicted-response decoding. Each dot denotes one cognitive label. Blue and red dots are derived from web and TV ad movie sets, respectively. The correlation is significant (Pearson r = 0.47; P < 0.0001), indicating that labels with higher decoding accuracy show higher individual-difference reflection.

S3 Supplementary Tables

Table S1: The number of ad movies belonging to individual product/service categories.

Categories	Web ad movies	TV ad movies
Electronic & Precision	4	9
Audiovisual	5	1
Appliance	16	5
Car	31	33
Food & Confectionery	7	69
Beverage & Alcoholic drink	20	38
Medical & Health	35	21
Cosmetics	49	6
Sundries & Home equipment	10	45
Garment/apparel	9	11
Entertainment	42	43
Media & Education	41	16
Distribution & Retailer	12	20
Communication & Service	35	51
House & Construction	9	12
Finance	9	20
Enterprise, Public service, & Others	34	20

Table S2: All items of cognitive labels associated with each movie set.

Categories	Items	Web ad movies	TV ad movies
Scene descriptions	-	✓	✓
Impression ratings	Beautiful	✓	✓
	Ugly	✓	✓
	Cute	✓	✓
	Nauseating	✓	✓
	Urban	✓	✓
	Rural	✓	✓
	Modern	✓	✓
	Traditional	✓	✓
	Human	✓	✓
	Mechanical	✓	✓
	Feminine	✓	✓
	Masculine	✓	✓
	Lush	✓	✓
	Cheap	✓	✓
	Intelligent	✓	✓
	Stupid	✓	✓
	Complex	✓	✓
	Simple	✓	✓
	Amusing	✓	✓
	Gloomy	✓	✓
	Quiet	✓	✓
	Noisy	✓	✓
	Dynamic	✓	✓
	Static	✓	✓
	Clean	✓	✓
	Filthy	✓	✓
	Warm	✓	✓
	Cool	✓	✓
	Bold	✓	✓
	Sensitive	✓	✓
Ad effectiveness indices	Click through rate	✓	
	25% view completion rate	✓	
	50% view completion rate	✓	
	75% view completion rate	✓	
	100% view completion rate	✓	
Ad preference votes	Preference		✓
	Cast/Character		✓
	Humorous		✓
	Sexy		✓
	Catchphrase		✓
	Music/Sound		✓
	Product attractiveness		✓
	Persuasive		✓
	Lame but lovable		✓
	Cutting-edge		✓
	Soothing		✓
	Story		✓
	Honest		✓
	Movie/Image		✓
	Reputable		✓
	Cute		✓
	Usage intention		✓
	Persistent use		✓
	Purchase intention		✓
Subjective preference ratings	-		✓

Table S3: Individual-difference reflection for each item of cognitive labels (*P < 0.05, **P < 0.01, ***P < 0.0001, FDR corrected).

		Individual-difference reflection	
Categories	Items	Web ad movies	TV ad movies
Scene descriptions	-	0.739***	0.720***
Impression ratings	Beautiful	0.354***	0.332***
	Ugly	0.271***	0.158**
	Cute	0.457***	0.399***
	Nauseating	0.303***	0.369***
	Urban	0.455***	0.388***
	Rural	0.405***	0.473***
	Modern	0.223***	0.290***
	Traditional	0.218***	0.097
	Human	0.394***	0.469***
	Mechanical	0.465***	0.253***
	Feminine	0.371***	0.302***
	Masculine	0.265***	0.151**
	Lush	0.275***	0.413***
	Cheap	0.368***	0.283***
	Intelligent	0.369***	0.421***
	Stupid	0.109**	0.440***
	Complex	0.179***	0.488***
	Simple	0.338***	0.311***
	Amusing	0.343***	0.361***
	Gloomy	0.235***	0.022
	Quiet	0.238***	0.551***
	Noisy	0.289***	0.052
	Dynamic	0.420***	0.367***
	Static	0.348***	0.329***
	Clean	0.428***	0.379***
	Filthy	0.108**	0.209***
	Warm	0.480***	0.264***
	Cool	0.390***	0.391***
	Bold	0.388***	0.548***
	Sensitive	0.438***	0.430***
Add effectiveness indices	Click through rate	0.157***	
	25% view completion rate	0.121**	
	50% view completion rate	0.193***	
	75% view completion rate	0.096**	
	100% view completion rate	0.112**	
Ad preference votes	Preference		0.204***
	Cast/Character		0.262***
	Humorous		0.287***
	Sexy		0.179**
	Catchphrase		0.115*
	Music/Sound		0.390***
	Product attractiveness		0.078
	Persuasive		0.112*
	Lame but lovable		0.106*
	Cutting-edge		0.181**
	Soothing		0.175**
	Story		0.176**
	Honest		0.092
	Movie/Image		0.370***
	Reputable		0.070
	Cute		0.234***
	Usage intention		0.156**
	Persistent use		0.198**
	Purchase intention		0.327***
Subjective preference ratings	-		0.395**