Disentangled deep generative models reveal coding principles of the human face processing network

Paul Soulos and Leyla Isik
Department of Cognitive Science, Johns Hopkins University

Abstract

Despite decades of research, much is still unknown about the computations carried out in the human face processing network. Recently deep networks have been proposed as a computational account of human visual processing, but while they provide a good match to neural data throughout visual cortex, they lack interpretability. We introduce a method for interpreting brain activity using a new class of deep generative models, disentangled representation learning models, which learn a low-dimensional latent space that "disentangles" different semantically meaningful dimensions of faces, such as rotation, lighting, or hairstyle, in an unsupervised manner by enforcing statistical independence between dimensions. We find that the majority of our model's learned latent dimensions are interpretable by human raters. Further, these latent dimensions serve as a good encoding model for human fMRI data. We next investigated the representation of different latent dimensions across face-selective voxels. We find a gradient from low- to high-level face feature representations along posterior to anterior face-selective regions, corroborating prior models of human face recognition. Interestingly, though, we find no spatial segregation between identity-relevant and irrelevant face features. Finally, we provide new insight into the few "entangled" (uninterpretable) dimensions in our model by showing that they match responses across the ventral stream and carry significant information about facial identity. Disentangled face encoding models provide an exciting alternative to standard "black box" deep learning approaches for modeling and interpreting human brain data.

Introduction

Humans are highly skilled at recognizing faces despite the complex high dimensional space that face stimuli occupy and the many transformations they undergo. Some dimensions (such as 3D rotation and lighting) are constantly changing and thus irrelevant to recognizing a face, while others (such as facial features or skin tone) are generally stable and useful for recognizing an individual's identity, and still others (such as hair style) can change but also offer important clues to identity. Face processing networks in the macaque and human brain have been thoroughly mapped [1]–[3] and many general coding principles have been identified, including separation of static vs. dynamic face representations [4], [5] and increasing transformation invariance from posterior to anterior regions [6]. However, much is still unknown about the computations carried out across these regions, particularly in the human brain. Even fundamental information, such as where and how facial identity is represented, is still largely unknown [2]. This lack of understanding can be seen in the relatively poor decoding of face identity from fMRI data compared to other visual categories [7].

Recently, deep convolutional neural networks (DCNNs) trained on face recognition have been shown to learn effective face representations that provide a good match to human behavior [8], but such discriminatively trained models are difficult to interpret [9] and provide a poor match to

human neural data [10]. Alternatively, deep generative models have been shown to provide a good match to human fMRI face processing data [11]. These models, however, transform faces into complex high dimensional latent spaces and thus suffer from the same lack of interpretability as standard DCNNs. Here we use a new class of deep generative models, disentangled representation learning models that isolate semantically meaningful factors of variation in individual latent dimensions, to understand the neural computations underlying human face processing.

Multiple disentangled representation learning models have been developed [12]–[17], many of which are based on Variational Autoencoders (VAEs) [18]. These disentangled variational autoencoders (dVAEs) learn a latent space that "disentangles" different explanatory factors in the training distribution by enforcing statistical independence between latent dimensions during training [19]. Intriguingly, when applied to faces, dVAEs have been shown to learn latent dimensions that are not only statistically independent, but also isolate specific, interpretable face features.

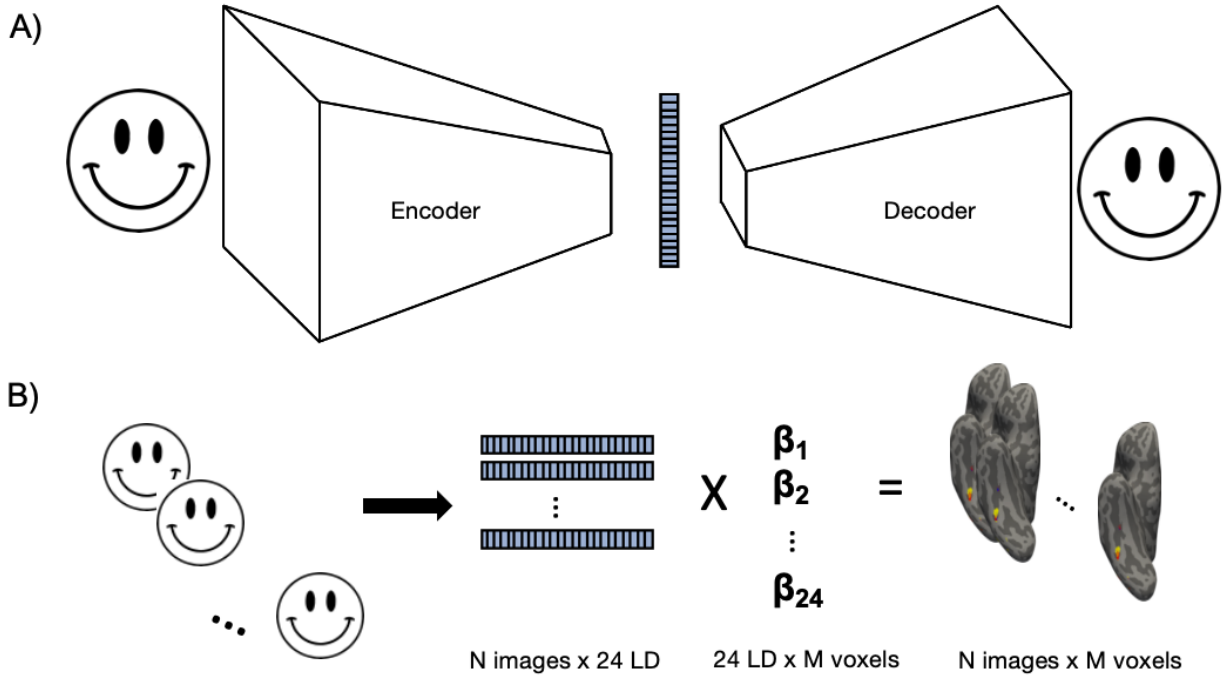

Fig 1: Model overview and encoding procedure. A) dVAE model overview. The model takes in a face image (left) and passes it through an encoder, consisting of several convolutional layers (left white trapezoid), to generate a latent vector (blue). Then a decoder (right white trapezoid) uses de-convolution to reconstruct the face image from the latent vector. Like a standard VAE, the dVAE is trained with a cost function to minimize reconstruction error. The dVAE has an additional term to maximize statistical independence (KL-divergence) between elements in the latent vector. B) Encoding procedure. A GLM is used to learn a linear mapping (beta weights) between the latent response to each training image and fMRI responses. At testing, a new test image is passed through the model to extract its 24-dimensional latent vector. This vector is then multiplied by the learned beta weights to generate a predicted voxel response.
Face images removed to comply with biorxiv face photograph policy. Unaltered figure available: http://bit.ly/3Iqv3Sg

dVAEs learn a latent representation that is compact and highly interpretable by humans, so we investigate complex face representations across the human brain using dVAEs in an encoding model framework. We find that representations in disentangled models match those found in the human face processing network at least as well or better than standard deep learning models without the disentanglement cost (including traditional VAEs and DCNNs). We then map the learned semantically meaningful dVAE dimensions to voxel responses and quantify their facial identity information, providing new insight into the models and the human face processing network.

Results

Disentangled generative models factor latent space into human-interpretable dimensions

We trained several dVAE models on the CelebA dataset [20], with the goal of selecting one as an encoding model of face-selective responses in the human brain. Like standard VAEs, these models have an encoder, which transforms an image into a lower-dimensional latent space via convolution, and a decoder, which aims to reconstruct the image from the latent representation (Fig. 1A). The models were trained to minimize reconstruction error and had an additional training objective to maximize KL divergence between latent dimensions. Based on a hyperparameter search to maximize disentanglement (see Methods M1), we selected FactorVAE [17] with 24 latent dimensions as our dVAE model.

After training the dVAE, two raters inspected the faces generated by traversing values of a single latent dimension while keeping all others constant. These latent traversals were often highly interpretable, producing faces that seem to vary along a single dimension, such as facial expression or 3D rotation (Fig. 2, Video S1-2). Out of the 24 latent dimensions, the human raters agreed on semantic labels for 16 (Table 1), which included both identity-relevant (dimensions 8-12, 14-16) and irrelevant (dimensions 1-7, 13). The other 8 were considered entangled, containing multiple or uninterpretable transformations.

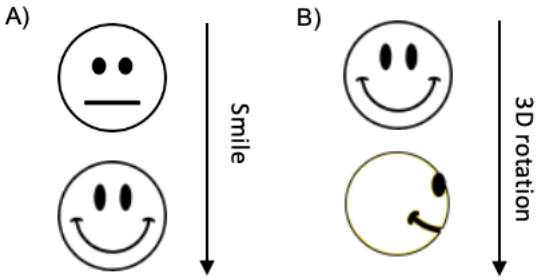

Fig 2: Latent traversals for two dVAE dimensions. Images generated by the dVAE when one latent dimension labeled as 'smile' (A) or '3D rotation' (B) is varied from -1 to +1, and other dimensions are held constant. The primary change in the image corresponds to the labeled dimension, suggesting these dimensions are effectively disentangled and highly interpretable.
Face images removed to comply with biorxiv face photograph policy. Unaltered figure available: http://bit.ly/3RYqIsm

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17-24
Lighting	Image Tone	Background	Background	3D rotation/ lighting	3D rotation	Head elevation	Hair part	Hair	Hair	Hair	Hairline	Smile	Skin tone	Gender appearance	Face Width	Entangled

Table 1: Labels assigned to 24 dimensions by human annotators. Annotators agreed on labels for 16 of 24 dimensions. Colors correspond to plots in Figs 5-6.

We compared our dVAE to two control models. First, we used a standard entangled generative VAE matched in terms of training and hyper-parameters. Second, we used the penultimate layer of a popular DCNN, the discriminatively trained VGG-Face based on VGG16 [21], [22]. To match model dimensions, we reduced the dimensionality of the VGG-Face representations to the first 24 principal components, which captured 70.7% of the variance. While the dVAE and VAE latent dimensions were highly correlated (CCA r = 0.92), the dVAE and VGG were only moderately correlated (CCA r = 0.52), suggesting that discriminative versus generative training frameworks result in different face representations.

Disentangled models provide a good match to ventral face-selective regions

We used a publicly available fMRI dataset [11], where four subjects viewed roughly 8000 face images each a single time. Each subject also viewed 20 face test images between 40-60 times. Data were pre-processed and projected onto subjects' individual cortical surfaces. We estimated a linear map between the latent representation of each model and the fMRI data via a generalized linear model (GLM) on the training data (Fig. 1b). To predict fMRI responses to each held out test image, we extracted the latent representation for that test image from each model and multiplied them by the linear mapping learned in the GLM.

We evaluated encoding performance for three face-selective ROIs, the Fusiform Face Area (FFA), Occipital Face Area (OFA), and posterior Superior Temporal Sulcus (pSTS), as well as face-selective voxels across the whole brain, identified in a separate face-object localizer experiment (see Methods). Despite the additional disentanglement constraint, the dVAE model achieves similar encoding performance to the standard VAE and VGG in FFA and OFA (Fig. 3, Table S1). At the group level, all models perform significantly above chance ($p<0.001$) in OFA and FFA. Additionally, both dVAE and VAE have significantly higher predictivity than VGG in the OFA and FFA at the group level (Table S1). The models also performed similarly across all face-selective voxels in the brain (Fig. 4, S1, and S2). None of the models provided consistently above chance accuracy in pSTS (Fig. 3), perhaps due to the fact that all stimuli were static faces and lateral face regions have been shown to be selective for dynamic stimuli [4].

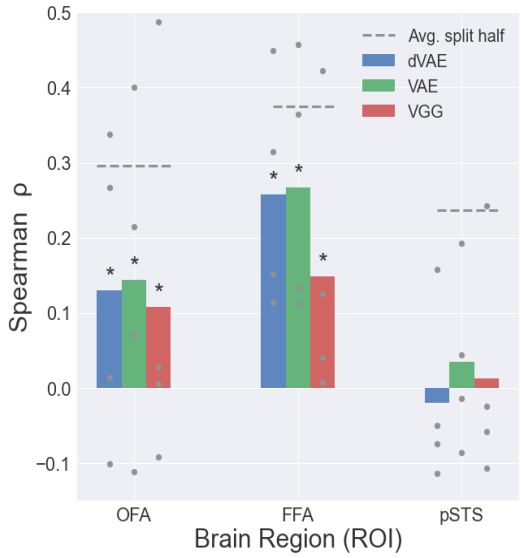

Fig 3: Encoding performance by ROI. Average correlation between model prediction and true fMRI responses on held out test images for dVAE (blue), VAE (green), and VGG (red). Gray dots represent individual subject performance. Asterisks represent significant ($p<0.05$) results at the group level based on permutation test. Dashed line is average split half reliability across subjects on test image responses.

Higher-level, identity-relevant dimensions are represented in more anterior face-selective regions

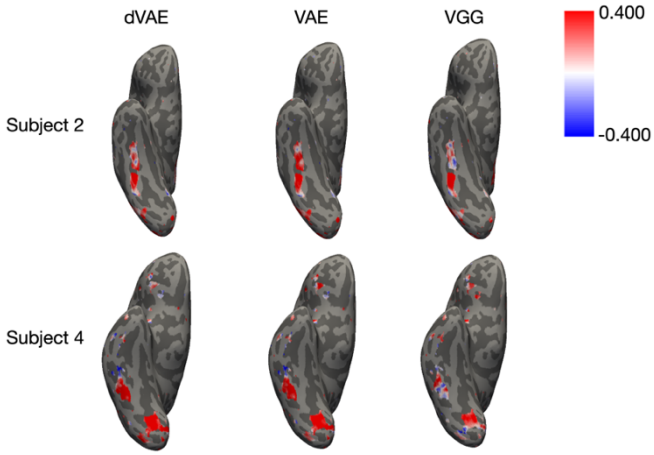

Fig 4: Whole brain encoding. Encoding model performance in all face-selective voxels. Ventral view of two representative subjects for dVAE (left), VAE (center) and VGG-face (right).

The main advantage of disentangled encoding models is the ability to examine how voxels respond to semantically meaningful dimensions. To do this, we performed preference mapping by predicting fMRI responses based on the dVAE latent vector and learned beta weights for each individual latent dimension. Preference mapping is similar to directly comparing the learned beta weights for each feature, but more robust since it is done on held out test data, and more interpretable since the outputs are bounded correlation values versus arbitrarily scaled beta weights [23]. High predictivity of a particular dimension in a particular brain region

indicates that changes along that dimension predict changes in neural activity and does not necessarily mean that specific region codes for or is selective to that dimension.

We first performed preference mapping within each ROI (Fig. 5). In the OFA, lower-level visual dimensions like lighting, image tone, and skin tone were significantly predictive in all subjects. Another visual dimension, background, as well as one entangled dimension, were predictive in three of four subjects. Like in OFA, background and skin tone also provided significant prediction in each subject in the FFA. Additionally, higher-level face-specific dimensions like smile, hair, 3D rotation, and entangled dimensions were also significantly predictive of FFA voxel responses in at least three subjects. These FFA-dimensions included both identity-specific features like skin tone and hair, and changeable aspects of faces like expression and 3D rotation. As with the full model performance in the STS, the performance from most individual dimensions is also worse than the other ROIs (Fig. S3).

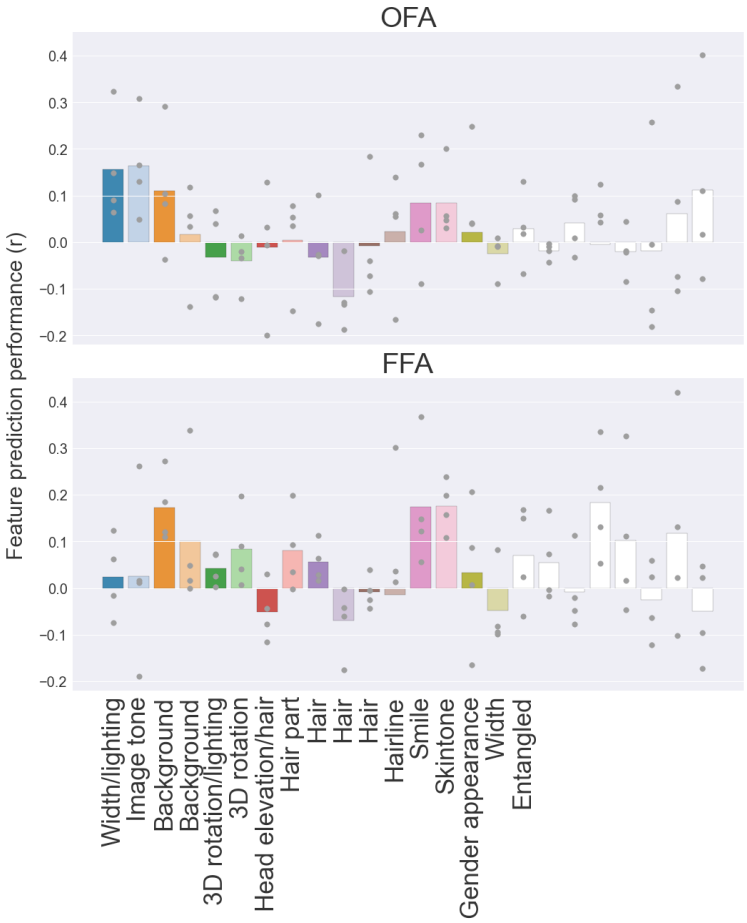

Fig 5: Preference mapping results for each dimension and ROI. Average feature predictivity across subjects is shown for each individual dimension. Gray dots represent individual subjects. Dimensions are colored according to Table 1.

To understand how dimensions are represented across the brain, we can visualize their predictivity in a winner take all manner on the surface of the brain (Fig. S4). Similar to the ROI analysis, most posterior voxels were best predicted by image-level changes in background and image tone. More anterior regions, including FFA, and in some subjects, anterior temporal lobe (ATL) regions not included in our ROI analysis, also showed responses for face-specific dimensions like smile (dark pink) and identity-relevant dimensions like skin tone (light pink), hairstyle (purples and browns), and gender appearance (green). Some subjects also show anterior ventral voxels best predicted by visual features like background (oranges). Interestingly, entangled dimensions (white) were predictive throughout the face-selective hierarchy.

Disentangled models isolate identity relevant face information

Another benefit of disentangled encoding models is the ability to study and group dimensions based on semantically meaningful attributes. One particularly important distinction for face processing is the separation of identity relevant factors (e.g., gender appearance, skin tone, and face shape) from identity irrelevant factors (e.g. lighting, viewpoint, and background). We decoded identity from our 20 test images using different subsets of dimensions: identity-relevant, identity-irrelevant, and entangled. Identity-relevant dimensions provided the highest identity decoding accuracy, almost equal to using all dimensions, whereas identity-irrelevant dimensions had the lowest, providing proof of concept that distinctions between our disentangled dimensions capture meaningful semantic information (Fig. 6). The role of information contained in entangled dimensions is an open question in AI, so we sought to examine the extent of identity information in these dimensions. The entangled dimensions contained some identity information as illustrated by their above-chance decoding. However, entangled dimensions do not appear to capture information beyond the identity-relevant dimensions as shown by the similar decoding performances of identity-relevant features and he combination of identity-relevant and entangled features.

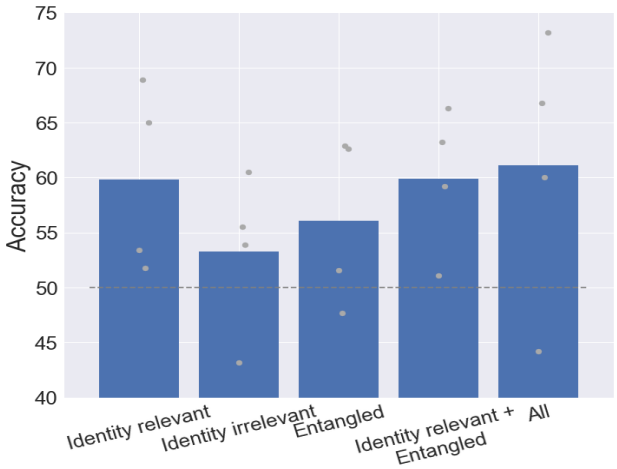

Fig 6: Facial identity decoding. Pairwise decoding of face identity from different subsets of dimensions: 8 identity relevant dimensions, 8 identity irrelevant dimensions, 8 entangled dimensions, the combination of identity relevant with entangled dimensions, as well as all dimensions. Mean values shown as blue bars with individual subjects shown as gray dots.

Discussion

We introduced a novel encoding framework for interpreting human fMRI data. Our method allows us to identify semantically meaningful dimensions in an unsupervised manner from large datasets. This disentanglement improves interpretability without a large degradation in encoding performance. Our results identified a gradient of low- to high-level properties represented along posterior to anterior brain regions, consistent with prior data and models of face processing [2], [24]–[26]. While we identified several identity-relevant dimensions in FFA, consistent with prior work [26], [27], we also found sensitivity to several changeable aspects of faces, including expression, in FFA and other ventral face-selective voxels. These results challenge the idea of a clear-cut distinction between identity and expression coding in ventral and lateral face regions [28]–[30]. In addition to mapping representations of disentangled dimensions across the brain, our approach also allows us to investigate the content contained in entangled dimensions. We found that entangled dimensions are represented throughout the face processing network, suggesting they code for both low- and high-level face properties. We also showed for the first time that the entangled dimensions contain identity-relevant information, providing new insight into their computational role.

Prior work has found that DCNNs trained for facial identity discrimination only capture a small amount of variance in human face selective regions [10] and do not replicate activity in the primate face patch hierarchy or human behavioral responses [31]. We see an advantage of our generatively trained encoding models versus the discriminatively trained DCNN particularly in the FFA, although this is not significant in all individual subjects (Fig. 3). The original paper presenting this fMRI dataset also found good decoding performance across the brain with a generative VAE, achieving much higher decoding performance than the results presented here [11], as have other studies comparing generatively trained neural networks to visual brain responses [32]–[34]. The original study focused on maximizing fMRI reconstruction and decoding with a high dimensional network (1024 dimensions vs. our 24). We chose our model to have the highest disentanglement which yielded the lowest dimensional network from all those tested in our hyperparameter search (see Methods M1), likely because enforcing statistical independence between latent dimensions via regularization during training becomes less effective as dimensionality increases. Thus our 24-dimensional network is less expressive than higher dimensional networks because it has a much smaller bottleneck for modeling the data distribution. However, the added interpretability afforded by disentanglement allows more fine-grained interpretability of the fMRI data not possible with standard models. Future work should investigate how to combine the interpretability benefits of disentangled models with the expressiveness of high dimensional networks.

Another recent approach has sought to learn disentangled latent representations in a supervised manner [31]. They learn a model which inverts a 3D face graphics program by supervising intermediate representations to match the primitives defined in the program (e.g. 3D shape, texture, and lighting). They find that this network matches primate face representations better than identity trained networks. Importantly, these intermediate representations are prespecified and need to be learned from labeled synthetic data. Many of these prespecified dimensions match those learned by our dVAE, providing further support for disentangled learning as a method to learn relevant latent dimensions in an unsupervised manner.

One recent prior study has investigated the correspondence between dVAEs and single neurons in macaque IT [35]. They find several IT neurons that show high one-to-one match with single units in their dVAE. They also demonstrate a high degree of disentanglement in the macaque neurons by showing a strong correlation between model disentanglement and alignment with IT neurons. It is worth noting that only a handful of neurons in their data show high alignment with single disentangled dimensions. Perhaps unsurprisingly given the lower spatial resolution of fMRI, we do not see the same high disentanglement in our data as evidenced by the fact that each region is well predicted by multiple latent dimensions. It remains an open question whether the primate face network is disentangled or shows exact correspondence to the dimensions learned by dVAEs at larger scale.

The content of the disentangled dimensions learned by our dVAE, and all other disentangled models, reflects the distribution of features in its training set. CelebA is a dataset of celebrity images which does not reflect the underlying distribution of faces that people see in daily life. In particular, CelebA faces tend to be young adults, white, and smiling at a camera. One example of how this can affect learned representations can be seen in the smile dimension, which is sometimes entangled with wearing sunglasses (Supplementary Video 1), likely reflecting a bias in CelebA that people wearing sunglasses tend to be smiling. More critically, the visual as well as racial and

ethnic biases in the dataset likely impact the quality of the learned dimensions [36]. Training models on a more ecologically valid dataset may improve encoding performance by better reflecting the statistics of real-world visual experience.

This work has important applications for cognitive neuroscientists to understand the relationship between semantic factors and neural activity using natural datasets without labels, in a scalable manner. As the quality of models and fMRI data increase, our method can be used to identify new semantically meaningful data dimensions with higher precision. Disentangled models have been created for various visual domains including object and scene processing [37]–[39] and can in theory be applied to any large scale visual dataset. While disentangled models are an active area of research in AI, there has been little investigation of their cognitive and neural plausibility. Our work sheds light on the role of entangled and disentangled dimensions in face representations in the brain and provides avenues for follow-up questions pertaining to their role in identity decoding. Understanding the neural coding of disentangled dimensions in the brain can help inspire novel data representations in AI.

Methods

M1. Neural Net Architecture and Training

We trained our VAE models using the TensorFlow DisentanglementLib package [40]. To identify the best disentangled model for our fMRI analyses, we performed a hyperparameter search over model architectures (including beta-VAE [14] and FactorVAE [17]), number of latent dimensions (24, 32, 48, and 64), and architecture-specific disentanglement parameters (beta-VAE $\beta \in [1, 2, 4, 6, 8, 16]$, FactorVAE $\gamma \in [10, 20, 30, 40, 50, 100]$). For every hyperparameter combination, we performed 10 random initializations. This resulted in 240 FactorVAE models and 240 beta-VAE models. We used beta-VAE without disentanglement (beta=1) for the standard, non-disentangled VAE models. After training, models were evaluated using the unsupervised disentanglement metric (UDR) [41]. We selected the model with the highest disentanglement score, a FactorVAE model with 24 latent dimensions and γ=10, for subsequent encoding analyses. Of the dimension-matched standard VAE models, we selected the randomly initialized model with the highest disentanglement score as our baseline.

For our baseline discriminative model, we used VGG-Faces [22], a network that uses the VGG architecture [21] and is trained from scratch on 2.6 million face images to predict face identity. To facilitate model comparison, we take the representations at the final fully connected layer and use Principal Component Analysis to reduce the dimensionality to match that of the VAEs.

M2. Dimension Annotation

After training and selecting our disentangled model, we passed 20 face images, not included in training, to the model. For each face image, we generated a set of "traversal images" by changing the value of a single latent dimension (e.g., Fig. 2) from -2 to +2. The traversal images for each latent dimension were combined into an animated gif and shown to two annotators. Annotators labeled each dimension in each image. We first consolidated the annotations for each annotator across images for each dimension by finding labels that were consistent across at least one third of the 15 face images. We then selected labels where both annotators agreed on the majority of images for our final labels (Table 1). The annotators agreed on 15 out of the 16 labeled dimensions.

For the one dimension that the annotators did not agree on, we included both labels (3D rotation/lighting). The remaining 8 dimensions were either not labeled or were not labeled consistently.

M3. fMRI Data and Preprocessing

We used publicly available fMRI data of four subjects from [11]. Subjects viewed around 8000 "training" face images each presented once, and 20 "test" face images presented between 40-60 times each. Face images were selected at random from the CelebA dataset and passed through a VAE-GAN. Each face was on the screen for 1s followed by a 2s ISI. The experiment was split over eight scan sessions. Subjects were also scanned on 8-10 separate face-object localizer runs to identify face-selective voxels. Data were collected on a Philips 3T ACHEIVA scanner. Subjects provided informed consent and all experiments were conducted in accordance with Comité de Protection des Personnes standards. For more details, refer to the original paper.

Data were pre-processed and projected onto subjects' individual cortical surfaces using Freesurfer [42]. Preprocessing consisted of motion correcting each functional run, aligning it to each subject's anatomical volume and then resampling to each subject's high-density surface. After alignment, data were smoothed using a 5 mm FWHM Gaussian kernel. All individual analyses were performed on each subject's native surface.

M4. ROI Definition

Regions of interest were defined using a group-constrained subject-specific approach [43]. The regions we investigated were the right Fusiform Face Area (FFA), Occipital Face Area (OFA), and Superior Temporal Sulcus (STS). To define our regions of interest (ROIs), we used the published group parcels from [43].

We selected the top 10% of voxels in each parcel using a metric that combined both face-selectivity and reliability on the test data. We first calculated face-selectivity based on face-object localizer runs, and z-scored each subjects' face>object p-values within each parcel to yield a selectivity score v_s for each voxel. We next calculated the split-half reliability in our test data (Spearman r), and z-scored these values within each parcel to generate a reliability score v_r. We then summed the normalized selectivity and reliability scores to yield our final selection metric ($v = v_s + v_r$). We restricted our ROI analyses to the right hemisphere because of more selective face responses and increased reliability in our test data. Across the subjects, the FFA had roughly 170 voxels, the OFA 110 voxels, and the STS 170 voxels. For our whole brain analyses, we computed the above metric ($v = v_s + v_r$) for each cortical voxel. We then selected all voxels that scored more than 1.5 standard deviations above the mean.

M5. Encoding Model Procedure

We estimated a linear map between the latent dimensions in our models and the fMRI data via a generalized linear model (GLM), following the procedure in the original study [11]. Since each training face image was shown only once, the latent values for that image (rather than the image itself) were included as weighted regressors to increase reliability of the learned beta weights. The latent values for each training face image, the test faces, and a general face "bias" term were all

included as regressors, as well as well as nuisance regressors for linear drift removal and motion correction (x, y, z) per run.

To test the accuracy of the encoding model, we extracted the latent dimensions for each test image and multiplied this by the beta weights learned in the GLM and adding the above "bias" term to get a predicted voxel response to each test image (Fig. 1). This produced an estimated brain response for each test image. We then compared the predicted fMRI response in each voxel activity to the true voxel activity across all test images using Spearman correlation.

M6. Preference Mapping

To understand the contribution of each latent dimension to brain responses, we followed the same encoding model training procedure described above. In model testing, we then generated the voxel prediction using a single latent dimension value instead of all the latent dimension values and calculated the correlation between the single dimension's predictions and ground truth. We calculated the average prediction for each latent variable within each ROI (Fig 5). For whole brain analyses, we performed preference mapping [23], assigning each voxel's preference label as the dimension which yielded the highest prediction.

M7. Identity Decoding

To understand the identity-relevant information in different latent dimensions we performed identity decoding of our test images. To decode identity, we took the same learned betas (W) from the encoding training procedure and multiplied the test fMRI data (y) by its pseudo-inverse ($\hat{x} = W^{-1} * (y - b)$) where b is the face bias term. This generated a predicted set of latent dimensions for each test image. We correlated the predicted latent dimensions with the true test latent dimensions and one random foil to assess the pairwise accuracy of the decoding. We repeated this for different subsets of latent dimensions: all those labeled as identity-relevant (including hair as it offers important cues to identity and prior work has shown sensitivity in face-selective voxels [44]), identity-irrelevant, and entangled dimensions.

M8. Statistical Testing

As the underlying distribution of our data was unknown, we used non-parametric, resampling-based statistics. To evaluate whether each model achieved above chance performance, we generated null hypotheses by repeating the above prediction correlations with shuffled test image labels 1000 resample runs. We performed shuffling within subject, and then computed p-values for each individual as well as group-average prediction.

To compare models, we take the difference in prediction and compare this to a null distribution with shuffled model labels. We generated 1000 resample runs and calculate p-values for each two-tailed pairwise model comparison at the individual and group levels (Table S1).

M9. Analysis Code

The code for the analysis is available at https://github.com/psoulos/disentangle-faces.

Acknowledgements

We thank Michael Bonner for helpful discussions on this work, and Emalie McMahon and Raj Magesh for feedback on the manuscript.

References

[1] D. Y. Tsao, S. Moeller, and W. A. Freiwald, "Comparing face patch systems in macaques and humans," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 49, pp. 19514–19519, Dec. 2008.

[2] B. Duchaine and G. Yovel, "A Revised Neural Framework for Face Processing," *Annu. Rev. Vis. Sci.*, vol. 1, no. 1, pp. 393–416, Nov. 2015.

[3] W. Freiwald, B. Duchaine, and G. Yovel, "Face Processing Systems: From Neurons to Real-World Social Perception," *https://doi.org/10.1146/annurev-neuro-070815-013934*, vol. 39, pp. 325–346, Jul. 2016.

[4] D. Pitcher, D. D. Dilks, R. R. Saxe, C. Triantafyllou, and N. Kanwisher, "Differential selectivity for dynamic versus static information in face-selective cortical regions," *Neuroimage*, vol. 56, no. 4, pp. 2356–2363, Jun. 2011.

[5] P. Polosecki, S. Moeller, N. Schweers, L. M. Romanski, D. Y. Tsao, and W. A. Freiwald, "Faces in Motion: Selectivity of Macaque and Human Face Processing Areas for Dynamic Stimuli," *J. Neurosci.*, vol. 33, no. 29, pp. 11768–11773, Jul. 2013.

[6] W. A. Freiwald and D. Y. Tsao, "Functional compartmentalization and viewpoint generalization within the macaque face-processing system.," *Science*, vol. 330, no. 6005, pp. 845–51, Nov. 2010.

[7] N. Kriegeskorte, E. Formisano, B. Sorger, and R. Goebel, "Individual faces elicit distinct response patterns in human anterior temporal cortex," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 51, pp. 20600–20605, Dec. 2007.

[8] A. J. O'Toole, C. D. Castillo, C. J. Parde, M. Q. Hill, and R. Chellappa, "Face Space Representations in Deep Convolutional Neural Networks," *Trends Cogn. Sci.*, vol. 22, no. 9, pp. 794–809, Sep. 2018.

[9] A. J. O'Toole and C. D. Castillo, "Face Recognition by Humans and Machines: Three Fundamental Advances from Deep Learning," *https://doi.org/10.1146/annurev-vision-093019-111701*, vol. 7, pp. 543–570, Sep. 2021.

[10] G. Jiahui *et al.*, "Modeling naturalistic face processing in humans with deep convolutional neural networks," *bioRxiv*, p. 2021.11.17.469009, Apr. 2022.

[11] R. VanRullen and L. Reddy, "Reconstructing faces from fMRI patterns using deep generative neural networks," *Commun. Biol.*, vol. 2, no. 1, p. 193, 2019.

[12] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep Convolutional Inverse Graphics Network," in *Advances in {Neural} {Information} {Processing} {Systems}*, 2015, vol. 28.

[13] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[14] I. Higgins *et al.*, "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," 2022.

[15] T. Carlson, D. A. Tovar, A. Alink, and N. Kriegeskorte, "Representational dynamics of object vision: the first 1000 ms.," *J. Vis.*, vol. 13, no. 10, pp. 1-, Jan. 2013.

[16] R. T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[17] H. Kim and A. Mnih, "Disentangling by Factorising," 2018.

[18] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," no. arXiv:1312.6114. May-2014.

[19] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[20] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild." pp. 3730–3738, 2015.

[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014.

[22] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC 2015 - Proceedings of the British Machine Vision Conference 2015*, 2015, pp. 1–12.

[23] H. Lee Masson and L. Isik, "Functional selectivity for social interaction perception in the human superior

temporal sulcus during natural viewing," *Neuroimage*, vol. 245, p. 118741, Dec. 2021.
[24] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini, "The distributed human neural system for face perception," *Trends Cogn. Sci.*, vol. 4, no. 6, pp. 223–233, Jun. 2000.
[25] B. Rossion, R. Caldara, M. Seghier, A. M. Schuller, F. Lazeyras, and E. Mayer, "A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing," *Brain*, vol. 126, no. 11, pp. 2381–2395, Nov. 2003.
[26] M. Tsantani, N. Kriegeskorte, K. Storrs, A. L. Williams, C. McGettigan, and L. Garrido, "FFA and OFA Encode Distinct Types of Face Identity Information," *J. Neurosci.*, vol. 41, no. 9, pp. 1952–1969, Mar. 2021.
[27] J. M. Contreras, M. R. Banaji, and J. P. Mitchell, "Multivoxel Patterns in Fusiform Face Area Differentiate Faces by Sex and Race," *PLoS One*, vol. 8, no. 7, p. e69684, Jul. 2013.
[28] J. Haxby, E. Hoffman, and M. Gobbini, "The distributed human neural system for face perception.," *Trends Cogn. Sci.*, vol. 4, no. 6, pp. 223–233, Jun. 2000.
[29] H. Zhang, S. Japee, R. Nolan, C. Chu, N. Liu, and L. G. Ungerleider, "Face-selective regions differ in their ability to classify facial expressions," *Neuroimage*, vol. 130, pp. 77–90, Apr. 2016.
[30] R. J. Harris, A. W. Young, and T. J. Andrews, "Brain regions involved in processing facial identity and expression are differentially selective for surface and edge information," *Neuroimage*, vol. 97, pp. 217–223, Aug. 2014.
[31] I. Yildirim, M. Belledonne, W. Freiwald, and J. Tenenbaum, "Efficient inverse graphics in biological face processing," *Sci. Adv.*, vol. 6, no. 10, p. eaax5979, Mar. 2020.
[32] K. Han *et al.*, "Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex," *Neuroimage*, vol. 198, pp. 125–136, Sep. 2019.
[33] 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*, vol. 181, pp. 775–785, Nov. 2018.
[34] Y. Güçlütürk, U. Güçlü, K. Seeliger, S. Bosch, R. van Lier, and M. A. van Gerven, "Reconstructing perceived faces from brain activations with deep adversarial neural decoding," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[35] I. Higgins *et al.*, "Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons," *Nat. Commun. 2021 121*, vol. 12, no. 1, pp. 1–14, Nov. 2021.
[36] A. Birhane and O. Guest, "Towards decolonising computational sciences."
[37] C. Lin, Z. Yuan, S. Zhao, P. Sun, C. Wang, and J. Cai, "Domain-Invariant Disentangled Network for Generalizable Object Detection." pp. 8771–8780, 2021.
[38] D. Epstein, T. Park, R. Zhang, E. Shechtman, and A. A. Efros, "BlobGAN: Spatially Disentangled Scene Representations," pp. 616–635, 2022.
[39] E. L. Denton and vighnesh Birodkar, "Unsupervised Learning of Disentangled Representations from Video," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[40] F. Locatello *et al.*, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *international conference on machine learning*, 2019, pp. 4114–4124.
[41] S. Duan *et al.*, "Unsupervised Model Selection for Variational Disentangled Representation Learning," 2019.
[42] B. Fischl, M. I. Sereno, and A. M. Dale, "Cortical Surface-Based Analysis," *Neuroimage*, vol. 9, no. 2, pp. 195–207, Feb. 1999.
[43] J. B. Julian, E. Fedorenko, J. Webster, and N. Kanwisher, "An algorithmic method for functionally defining regions of interest in the ventral visual pathway.," *Neuroimage*, vol. 60, no. 4, pp. 2357–64, May 2012.
[44] W. A. Freiwald, D. Y. Tsao, and M. S. Livingstone, "A face feature space in the macaque temporal lobe," *Nat. Neurosci.*, vol. 12, no. 9, pp. 1187–1196, Sep. 2009.

Supplementary Information

Factor vs VAE

	OFA	FFA	pSTS
Group	VAE (.001)	VAE (.009)	VAE (.001)
Subject 1	Factor (.166)	Factor (.001)	VAE (.001)
Subject 2	VAE (.001)	VAE (.001)	VAE (.001)
Subject 3	VAE (.001)	VAE (.005)	VAE (.001)
Subject 4	Factor (.001)	VAE (.103)	VAE (.001)

Factor vs VGG

	OFA	FFA	pSTS
Group	Factor (.035)	Factor (.001)	VGG (.001)
Subject 1	VGG (.001)	Factor (.001)	VGG (.059)
Subject 2	VGG (.001)	VGG (.001)	VGG (.003)
Subject 3	Factor (.001)	VGG (.099)	Factor (.01)
Subject 4	Factor (.001)	Factor (.001)	VGG (.001)

VAE vs VGG

	OFA	FFA	pSTS
Group	VAE (.001)	VAE (.001)	VAE (.022)
Subject 1	VGG (.001)	VAE (.001)	VAE (.236)
Subject 2	VGG (.001)	VGG (.001)	VGG (.084)
Subject 3	VAE (.001)	VAE (.128)	VAE (.001)
Subject 4	VAE (.001)	VAE (.001)	VGG (.001)

Table S1: Group and individual subject model versus model significance results. P-value are in parantheses.

Video S1-2: Animated latent traversals for all 24 latent dVAE dimensions for two example rendered faces. Dimensions are varied from -2 to +2, with all other dimensions held constant.
https://static.wixstatic.com/media/19a669_67da0720a32946568470b4b819c444eb~mv2.gif
https://static.wixstatic.com/media/19a669_5a3bc8a8f4654aa0b1051d254e56453c~mv2.gif

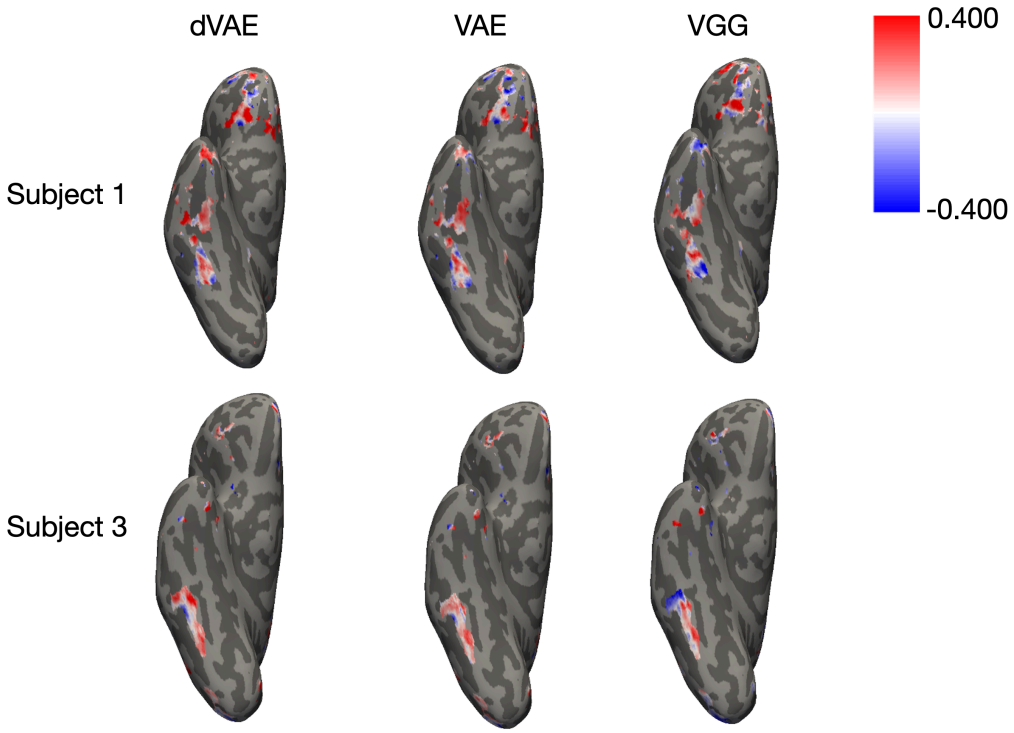

Fig S1: Whole brain encoding. Encoding model performance in all face-selective voxels. Ventral view of the remaining two subjects for dVAE (left), VAE (center) and VGG-face (right).

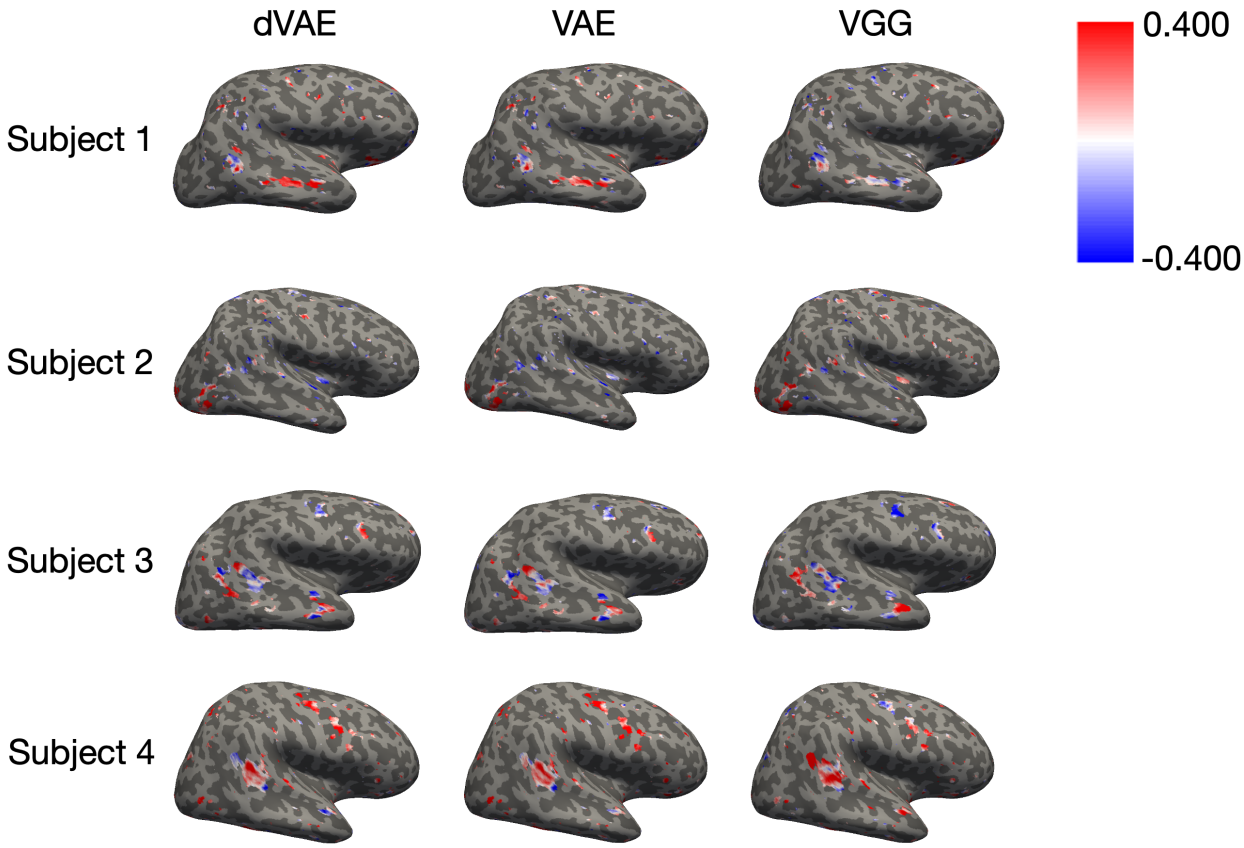

Fig S2: Whole brain correlations for all 4 subjects' lateral view.

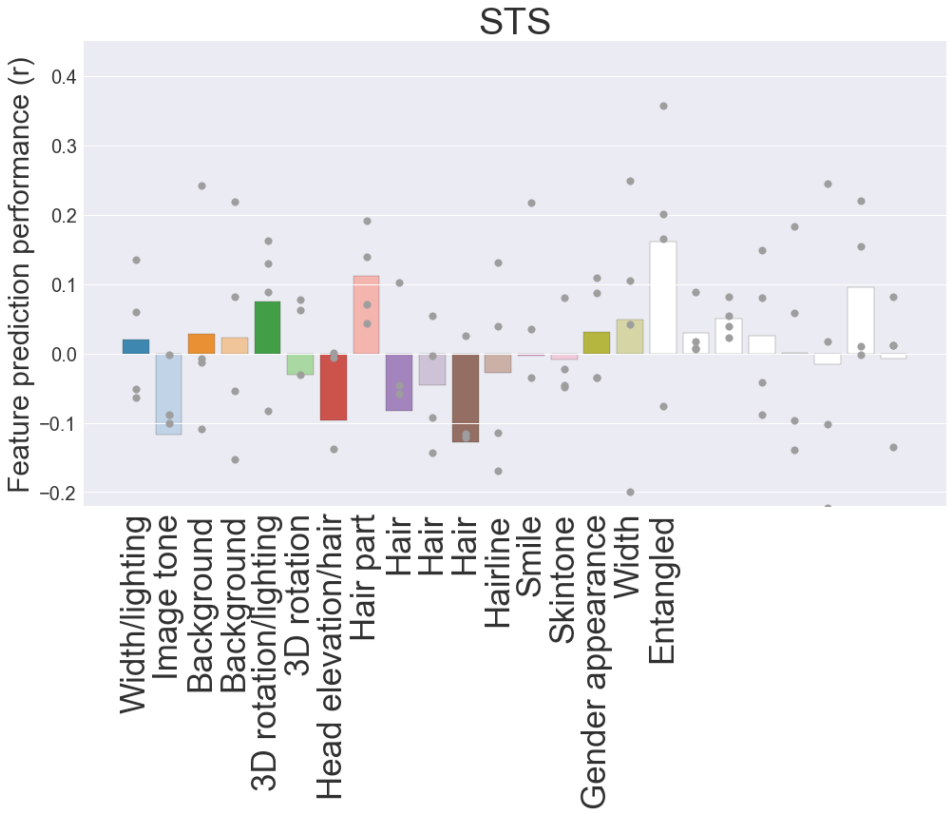

Fig S3: Preference mapping for each dimension in the STS. Hair part was the only dimension that provided significantly above chance in all subjects. 3D rotation and one entangled dimension were also significant in three of four subjects.

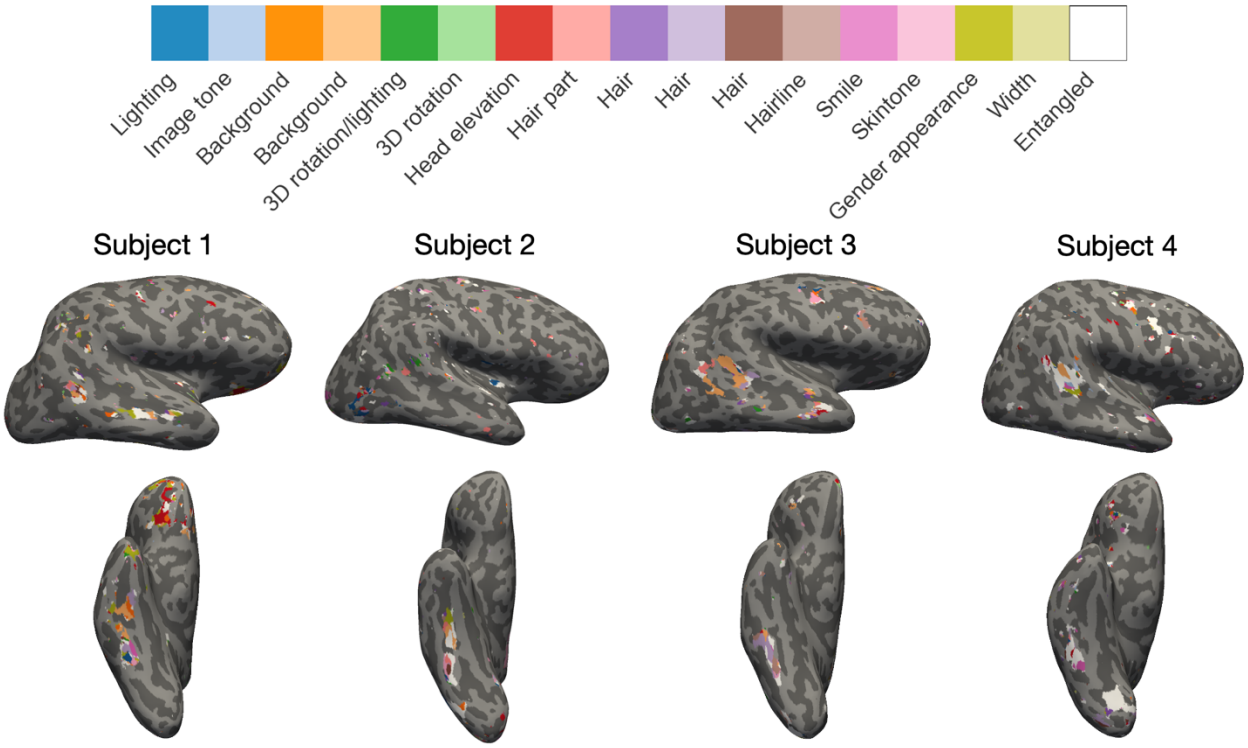

Fig S4: Whole brain preference maps. Each voxel is shaded based on the latent dimension that provides the highest predictivity.