JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles | Behavioral/Cognitive

Multidimensional dynamics of object representations in the human visual system

https://doi.org/10.1523/JNEUROSCI.0623-26.2026

Received: 10 April 2026
Revised: 31 October 2026
Accepted: 5 November 2026

Copyright © 2026 the authors

This Early Release article has been peer reviewed and accepted, but has not been through the composition and copyediting processes.The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

Multidimensional dynamics of object representations in the human visual system

Zirui Chen & **Leyla Isik** & **Michael F. Bonner**

Department of Cognitive Science

Johns Hopkins University, Baltimore, MD 21218

zchen160@jhu.edu, lisik@jhu.edu, mfbonner@jhu.edu

November 9, 2026

Abbreviated title: Multidimensional dynamics of object representation

Corresponding author: Zirui Chen, Department of Cognitive Science, Johns Hopkins University, Baltimore, MD 21218; zchen160@jhu.edu

Number of figures: 5 **Number of tables:** 0 **Number of multimedia:** 0 **Number of supplemental figures:** 5

Number of words – Abstract: 185 **Significance Statement:** 117 **Introduction:** 589 **Discussion:** 1315

Conflict of interest: The authors declare no competing financial interests.

JNeurosci Accepted Manuscript

Abstract

Natural image representations are distributed across many dimensions of visual cortex activity, but little is known about how the multidimensional structure of these representations evolves over time following stimulus onset. Here we examined the temporal dynamics and latent

dimensional structure of natural object representations in large-scale EEG and MEG data from male and female human participants. We also compared these data with leading representational models derived from large-scale human similarity judgments and deep neural networks. Our findings reveal a rapid expansion of stimulus dimensionality in the brain, which peaks within 100 milliseconds and gradually decays over hundreds of milliseconds. The dynamics of these dimensionality changes tracked the decoding accuracy for both behavioral embeddings and neural network features, suggesting that dimensionality may be a general indicator of representational expressivity. Interestingly, the dimensionality of the neural representations could not be fully explained by a leading behavior-based model or deep neural network. Follow-up experiments showed that the remaining neural variance carried additional perceptually relevant information not yet explained by leading models. Together, these findings reveal previously unrecognized complexity in measurements of dynamic human brain responses to natural objects.

Significance Statement

When we look at an object, our brains rapidly construct a rich internal representation of what we see. Using large-scale, time-resolved brain recordings from people viewing thousands of natural images, we find that the brain rapidly expands its representational complexity within the first 100 milliseconds, sustaining this richness for hundreds of milliseconds thereafter. A leading behavioral model and a large-scale deep neural network do not fully account for this complexity — yet the unexplained brain activity is not noise: people can reliably use these residual neural patterns to judge image similarity. These findings challenge the assumption that such brain recordings are informationally limited and reveal structure in the human visual system that future models will need to explain.

JNeurosci Accepted Manuscript

1 Introduction

How are natural image representations transformed along the processing stages of the human visual system? One perspective suggests that visual cortex implements dimensionality reduction, transforming high-dimensional sensory inputs into robust, low-dimensional representations of behaviorally relevant information (Fusi et al., 2016; Lehky et al., 2014; Lehky and Tanaka, 2016; Slapik and Shouval, 2026; Barrett and Miller, 2025). This view has motivated researchers to study visual representations by focusing on low-dimensional embeddings of neural activity (Connolly et al., 2012; Huth et al., 2012; Khosla et al., 2022; Thorat et al., 2019). However, advances in large-scale neural data collection (Allen et al., 2022; Hebart et al., 2023a; Chang et al., 2019) and analysis methods have shown that visual cortex representations of natural stimuli are not restricted to a low-dimensional subspace—instead they are scale-free, with stimulus-related information encoded along all available dimensions (Gauthaman et al., 2025; Stringer et al., 2019; Han and Bonner, 2026). These findings suggest that conventional approaches focusing on leading dimensions may capture only a small fraction of the meaningful information encoded in cortical activity.

Despite these advances, the dynamics of multidimensional structure in human visual representations remain largely unexplored. Many previous studies of dynamics in human vision have examined feature decoding in electroencephalography (EEG) and magnetoencephalography (MEG) data and identified representational features with distinct temporal profiles, with more complex features tending to emerge at later time points (Isik et al., 2014; Carlson et al., 2013; Cichy et al., 2014; Cichy et al., 2016; Coggan et al., 2016; Grootswagers et al., 2017; Teichmann et al., 2026; Greene and Hansen, 2018; Greene and Hansen, 2020; McMahon et al., 2025; Grootswagers et al., 2019). However, what is not known is how the dimensionality of these representations evolves over time and how it relates to the dynamics of feature decoding.

A challenge of analyzing the multidimensional structure of EEG and MEG data is that these recording techniques suffer from signal spread across channels and lower signal-to-noise

JNeurosci Accepted Manuscript

ratio compared with fMRI (Michel and Brunet, 2019; Grootswagers et al., 2017). Because of this, many EEG/MEG studies have employed dimensionality reduction as a preprocessing step (Bankson et al., 2018; Carlson et al., 2013; Hansen et al., 2019). However, in large-scale neural datasets, even dimensions with a low signal-to-noise ratio can carry reliable stimulus-related information and may be behaviorally relevant (Gauthaman et al., 2025; Han and Bonner, 2026). Thus, by restricting analyses to leading dimensions and pre-determined features, previous studies may have overlooked potentially rich structure in the remaining neural dimensions.

Here we characterized the multidimensional structure of stimulus-related signal in human EEG and MEG responses to natural images. We examined two large-scale datasets containing responses to thousands of object images, and we compared neural representations with model features derived from behavioral similarity judgments and deep neural networks (DNNs) (Gifford et al., 2022a; Hebart et al., 2023a; Hebart et al., 2020). Our analyses show that the representational dimensionality of human brain responses rapidly increases after stimulus onset and remains high for hundreds of milliseconds. Conventional decoding analyses show that features derived from behavioral embeddings and DNNs are represented most strongly at times of peak neural dimensionality. However, further analyses using cross-decomposition show that these leading representational models do not fully explain the dimensionality of the neural signal. A follow-up behavioral experiment demonstrates that the remaining unexplained variance in neural responses contains perceptually relevant information. Together, these findings show that large-scale EEG and MEG data contain rich multidimensional structure that evolves rapidly after stimulus onset, carries behaviorally relevant information, and extends beyond what these common representational models capture.

JNeurosci Accepted Manuscript

2 Materials and Methods

2.1 Datasets

2.1.1 Dynamic neural data

We examined neural responses to natural object images from the THINGS database, which contains images from 1,854 diverse object categories (Hebart et al., 2019). We analyzed two large-scale publicly available datasets of dynamic neural responses to these images. The EEG dataset ("THINGS-EEG2") includes ten subjects (8 female, 2 male; mean age 28.5 years) who viewed THINGS images and determined if a target image showed up in a given sequence (Gifford et al., 2022a). Each stimulus was presented for 100 ms, followed by 100 ms of fixation, after which the next stimulus was shown. The training set includes four repetitions of responses to ten unique images per 1,654 concepts; the test set includes 80 repetitions of responses to one image per 200 concepts, which do not overlap with the concepts in the training set. The epoched data range from 200 ms before stimulus onset to 800 ms after stimulus onset. Of the 63 recorded scalp electrodes, our main analyses used the 17 channels overlying occipital and parietal cortex (O1, Oz, O2, PO7, PO3, POz, PO4, PO8, P7, P5, P3, P1, Pz, P2, P4, P6, P8), selected and preprocessed by Gifford et al. (2022a).

The MEG dataset ("THINGS-MEG") (Hebart et al., 2023a) includes four subjects (2 female, 2 male; mean age 23.25 years) who viewed THINGS images and responded whenever they detected an oddball synthetic image. Each stimulus was presented for 500 ms, followed by 1 second fixation, with ±200 ms of jitter, after which the next stimulus was shown. This dataset includes images from 1,854 object concepts. For all concepts, at least 12 unique images were presented once, but for a subset of 200 concepts, an additional unique image was presented 12 times. We used these 200 concepts with 12 repeated presentations as our test set. We then created a training set using all concepts other than these 200 test concepts. Specifically, our training set consisted of 1,654 concepts, each with 12 unique images, yielding a total of 19,848 training images. This means that in both the EEG and MEG datasets, all test-set evaluations are performed on images from

JNeurosci Accepted Manuscript

held-out concepts that did not appear in the training set (as opposed to held-out exemplars from the same concept categories). We sought to match the preprocessing procedures across the EEG and MEG datasets as much as possible, so we performed our own preprocessing on the raw MEG data with a few differences relative to the preprocessing performed by the original authors. These differences include bandpass filtering the raw data for each run between 0.1 and 100 Hz, baseline correcting the epoched data by subtracting the mean of the data during baseline 100 ms before stimulus onset, and downsampling to 100 Hz. Because of differences in stimulus timing across the EEG and MEG experiments, we used the same epoch time window for the MEG data as the original authors (Hebart et al., 2023a), which ranged from –100 ms to 1300 ms relative to stimulus onset. Of the 271 usable MEG sensors, our main analyses used the 80 sensors overlying occipital and parietal cortex (sensors with the prefixes MLO, MRO, MLP, and MRP), chosen to match the occipital–parietal selection used for the EEG data.

2.1.2 Behavioral embeddings

We analyzed 66-dimensional sparse positive object similarity embeddings (SPoSE) for the 1,854 object categories. These embeddings were learned in previous work to explain approximately 4.7 million trials of human odd-one-out judgments on the THINGS images (Hebart et al., 2020; Hebart et al., 2023a). These dimensions span both low-level visual properties, such as color and shape, and more conceptual categories (Teichmann et al., 2026).

2.1.3 Neural network feature maps

We extracted feature maps from OpenCLIP ResNet50, a state-of-the-art vision neural network trained on the YFCC15M dataset (Radford et al., 2021; Ilharco et al., 2021). We extracted activations from the first and last residual units, using the Identity layer following the middle ReLU. We reduced the dimensionality of these DNN activations using sparse random projections, with the number of projections determined via the Johnson-Lindenstrauss lemma, resulting in 8,336 dimensions.

JNeurosci Accepted Manuscript

2.2 Linear mapping methods

2.2.1 Feature decoding

We used ridge regression to predict features of interest from neural responses at each time point. For all analyses, we used a ridge penalty of $\alpha = .01$. We confirmed that our findings are not strongly contingent on the specific value of the ridge penalty. For each subject and time point, regression weights were learned using the training set neural responses averaged across repetitions. Decoding scores were computed by taking the Pearson correlation between the predicted and actual feature values in the test set, with statistical significance assessed using cluster-based permutation tests (Maris and Oostenveld, 2007). When computing decoding scores for the temporal generalization analyses, the regression weights learned at a given time point were applied to the test-set neural responses at all other time points.

2.2.2 Neural dimensionality

We assessed the dimensionality of stimulus-related variance in neural responses at each time point (Stringer et al., 2019; Gauthaman et al., 2025; Han and Bonner, 2026). We first learned a set of principal component (PC) directions from the training-set responses averaged across all repetitions, X_{train}:

$$\text{cov}(X_{\text{train}}, X_{\text{train}}) = U\Sigma V^{\top} \tag{1}$$

and then evaluated how reliably these directions were expressed in responses to the held-out test images. Because each test image was presented repeatedly (see Datasets), we estimated this reliability across independent repetitions of the same stimuli: for each test stimulus we randomly split its repetitions into two halves and averaged within each half to obtain X^a_{test} and X^b_{test}, which we projected onto the training PCs as $X^a_{\text{test}}U$ and $X^b_{\text{test}}U$, centering both using the training-set mean. At each rank, reliability was the Pearson correlation between the two projections across

JNeurosci Accepted Manuscript

the test stimuli. We repeated this per-stimulus split ten times and averaged the reliability at each rank. A dimension was counted as significant if its correlation exceeded a null distribution assessed with cluster-based permutation tests (Maris and Oostenveld, 2007), and we summarized the representational dimensionality at each time point as the number of significant dimensions. For temporal generalization analyses, the eigenvectors learned at a given time point were applied to test-set neural responses at the same time point and at all other time points. Correlations were then computed across time points rather than across stimulus repetitions.

2.2.3 Shared-space dimensionality

We computed the shared dimensionality between neural responses and representational features using a similar approach to the neural dimensionality analyses. Given the training-set neural responses averaged across all repetitions, X_{train}, and the corresponding feature values, Y_{train}, we performed a singular value decomposition (SVD) of their cross-covariance:

$$\text{cov}(X_{\text{train}}, Y_{\text{train}}) = U\Sigma V^{\top} \tag{2}$$

We then projected the held-out test-set responses and features onto this shared space as $X_{\text{test}}U$ and $Y_{\text{test}}V$, centering both using the training-set means. To match the amount of data contributing to each estimate in the neural dimensionality analysis, we applied the same per-stimulus split of the test repetitions: for each of ten random splits we computed the rank-wise correlation for each of the two neural half-averages against the same feature projection, and averaged across halves and splits. Note that this correlation is computed between the neural and feature projections, rather than between two halves of the neural data as in the analysis above. We defined the shared dimensionality as the number of dimensions whose correlation was significantly above the null distribution assessed with cluster-based permutation tests (Maris and Oostenveld, 2007).

JNeurosci Accepted Manuscript

2.3 Behavioral experiment

We conducted an image matching experiment to assess if reliable neural signals unexplained by existing behavioral embeddings are also behaviorally meaningful. We chose to focus on the EEG data at time point 200 ms, approximately where the shared dimensionality between EEG data and behavioral embeddings peaks. We first trained a ridge regression ($\alpha = .01$) to predict the train-set neural data, averaged across all repetitions, from the behavioral embeddings. We then computed the residual variance in the test set by subtracting the predicted values from the real test-set responses, averaged across all repetitions. We calculated the pair-wise Pearson distance (one minus Pearson correlation coefficient) between the residual responses to the test stimuli for each subject and averaged these values across subjects.

Participants for both the main and replication experiments were recruited online through Prolific; their sex and age were not collected. In the experiment, participants ($N = 50$) were presented with a target image on top and two groups of five images below. They were asked which group the target image should belong to. The correct group contains five images with the closest Pearson distance to the target image, and the foil group contains five images that were randomly selected from images whose distance to the target was above a fixed threshold (we used the 0.8 quantile of all pairwise distances in the first experiment, and we relaxed this to the 0.65 quantile in the replication experiment on EEG responses at 100 ms). Experimental trials were created for all 200 test images. The correct image group appeared randomly on the left or right side of the display across trials. The participants first completed a practice trial with the correct group being all the same image as the target images. Then they proceeded to complete a sequence of 100 randomly sampled experiment trials and ten catch trials in the same format as the practice trial. Experimental trials were sampled such that across sets of two participants, all 200 trials would be shown (i.e., for every 100 trials selected for one participant, the remaining 100 would be shown to another participant).

JNeurosci Accepted Manuscript

3 Results

3.1 Time-resolved neural responses exhibit rich multidimensional structure

We examined the dynamics of representational dimensionality in human brain responses to natural images. To do so, we leveraged two large-scale datasets of time-resolved responses to thousands of images, THINGS-EEG2 and THINGS-MEG (Gifford et al., 2022a; Hebart et al., 2023a) (Fig. 1a). The scale and richness of these datasets make them ideally suited for characterizing the multidimensional structure of visually evoked EEG and MEG responses. Both datasets contain responses to images of objects in their natural contexts, and they were sampled from the THINGS database (Hebart et al., 2019).

In both studies, subjects viewed the images while maintaining central fixation and performing an orthogonal target-detection task to ensure they were paying attention. Besides differing in imaging methods, these datasets also notably differ in their stimulus presentation timing, with the EEG dataset having a much shorter stimulus presentation time (100 ms for EEG and 500 ms for MEG) and a much shorter inter-stimulus interval (100 ms for EEG and 800-1,200 ms for MEG). The stark differences in the imaging methods and experimental paradigms of these datasets allow us to reveal dimensionality patterns that are either dependent on or invariant to experiment settings.

We characterized the dimensionality of stimulus representations in these data by applying principal component analysis (PCA) and estimating the reliability of the principal components (PCs) in held-out test data (Gauthaman et al., 2025; Stringer et al., 2019; Han and Bonner, 2026). For each time point, we learned the PCs of multi-channel response patterns to all training images. We then projected the test-set responses onto these PC bases and evaluated the reliability of each PC by computing its correlation across split halves of the stimulus repetitions. We computed a summary statistic of representational dimensionality by counting the number of PCs with significant reliability ($p < 0.05$, cluster-based permutation test). We performed these analyses for all channels in the occipital and parietal cortices, which encompassed 17 channels in the EEG data and 80 channels

JNeurosci Accepted Manuscript

in the MEG data. Computing this summary statistic at each time point allowed us to evaluate how the dimensionality of stimulus representations evolves over time.

The results are illustrated in Figure 2. For both datasets, they reveal a rapid increase in dimensionality after stimulus onset, reaching peak levels within 100 ms in the EEG data and 150 ms in the MEG data. Peak dimensionality is then sustained for hundreds of milliseconds (until 300 ms in the EEG data and 400 ms in the MEG data), followed by a gradual decay extending to nearly 800 ms. The similarities across the two datasets are notable given the major differences in their experimental paradigms. In the MEG experiment, the stimuli were displayed for 500 ms with a long interstimulus interval, allowing for the representations to evolve after stimulus onset without interference from other image inputs. However, even when stimuli are presented for a brief window of 100 ms and followed by a rapid succession of other sensory inputs, as in the EEG experiment, representational dimensionality nonetheless maintains peak levels for hundreds of milliseconds before gradually decaying. This suggests that the dimensionality dynamics observed here reflect a general property of image processing in the brain and are not strongly contingent on the details of stimulus presentation timing. At peak, the number of significant dimensions averaged 16.4 of the 17 EEG channels and 24.8 of the 80 MEG sensors, so the EEG estimates approach the ceiling imposed by the number of recording channels while the MEG estimates remain well below it.

In a control analysis, we performed the same procedure in a region where we expected to find less stimulus-related signal and thus lower dimensionality (Fig. S1). Specifically, we examined the 21 channels from the frontal lobes in the EEG data and the 64 channels from the frontal lobes in the MEG data. As expected, the dimensionality of stimulus representations in the frontal lobes was low and transient. This contrasts with the findings in the occipital and parietal lobes, where the dimensionality was much higher and sustained for hundreds of milliseconds. These negative-control findings indicate that the results in Figure 2 specifically reflect the dimensionality of reliable stimulus-related signal.

Together, these findings show that natural images evoke rich multidimensional representations

JNeurosci Accepted Manuscript

that rapidly evolve after stimulus onset and maintain their peak dimensionality for hundreds of milliseconds despite other ongoing sensory inputs. These findings further demonstrate that such multidimensional representations can be detected in large-scale EEG and MEG data and are specific to posterior channels over visual cortex.

3.2 Feature decoding tracks dimensionality

How do the dimensionality trends observed in Figure 2 relate to more conventional analyses of feature decoding? To address this question, we examined the decoding of both behavior-derived and DNN features, which have been a focus of previous work on object representations in EEG and MEG data (Cichy et al., 2016; Teichmann et al., 2026). For the behavior-derived features, we examined a set of 66-dimensional embeddings that were learned in previous work to explain a large set of human similarity judgments on the objects from the THINGS database (Hebart et al., 2020; Hebart et al., 2023a). For the DNN features, we examined a ResNet50 model trained on a large-scale dataset (YFCC15M) with a language alignment objective. Such language-aligned DNNs have been shown to be leading representational models of human visual cortex (Doerig et al., 2025; Wang et al., 2023). We extracted activations from the first and last residual layers of this DNN to capture low- and high-level visual features.

For the decoding analyses, we fit linear regression models to predict behavioral and DNN features from EEG/MEG responses. For each time point, we fit these regression models using the training data and obtained decoding scores by computing the correlation between the actual and predicted feature values in held-out test data. Figure 2 shows the decoding score at each time point averaged across each set of behavioral and DNN features. Consistent with previous work (Cichy et al., 2016; Teichmann et al., 2026), each of these sets of features are significantly decodable within 100 ms after stimulus onset and remain significant for hundreds of additional milliseconds, while slowly decreasing in strength. As expected, we also observe different time courses for low- and high-level visual features. Specifically, we see that the decoding accuracy for early DNN features increases rapidly, reaching a peak around 100 ms, whereas the decoding accuracy for both behavioral and

JNeurosci Accepted Manuscript

late DNN features rises more gradually, reaching its maximum between approximately 200 and 400 ms. Analyses of other DNN layers are reported in Figure S2, and they show the expected temporal trends, with earlier layers tending to exhibit peak decoding accuracy at earlier time points.

When we compare the decoding results with the dimensionality trends in Figure 2a, we see that the strongest decoding accuracies are generally observed during the temporal window when neural dimensionality is sustained near its highest level. We also see a correspondence between the tails of the time courses for both decoding accuracy and dimensionality, which exhibit a similar pattern of slow decay. Together, these findings suggest that the dimensionality of stimulus-related neural activity may be a general indicator of expressivity, with times points of high dimensionality supporting the decoding of diverse behavioral and computational features. Furthermore, these findings suggest that neural dimensionality remains at its peak level even as the representations transform from low- to high-level features, as evidenced by the sustained dimensionality from approximately 100-400 ms despite ongoing changes in feature decoding.

3.3 Representations dynamically transform after stimulus onset

We next sought to understand how the underlying dimensions of these representations change over time. We began by selecting three example time points, and we visualized how the PCs from these time points evolve. As in the dimensionality analyses of Figure 2a, we learned the PC bases for each time point in the training data. However, instead of examining only the corresponding time point in the test data, we also examined cross-correlations with all other time points. As shown in Figure 3a-b, these analyses reveal that in both the EEG and MEG data, the PCs from early, middle, and late time points exhibit distinct trajectories, confirming that the representational content varies even during times when the overall dimensionality remains relatively stable. These analyses also show that the dynamics of representational changes are faster in the EEG data, which is consistent with the faster stimulus-presentation protocol in that study.

We next expanded these analyses to all time-by-time comparisons and generated matrices summariz-

JNeurosci Accepted Manuscript

ing the number of correlated dimensions for each comparison (Fig. 3c-d). When performing these analyses, we noticed that some cross-time comparisons in the EEG data exhibited predominantly negative correlations. To characterize this effect, we computed two dimensionality metrics: one for positive correlations and another for negative correlations. We used these two dimensionality metrics to generate the summary plots in Figure 3c-d, with each cell displaying the larger of the two metrics. Both the EEG and MEG data exhibit a band of high positive dimensionality along the diagonal, indicating that the multidimensional representations at each time point are reliable and at least partially distinct from other time points. However, the off-diagonal elements of these matrices show a clear difference across the EEG and MEG data. Specifically, in the EEG results, the dimensionality score fluctuates from large positive to large negative values, suggesting that all the dimensions of the representation are oscillating. We wondered if these oscillations might be related to the stimulus-presentation rate. To investigate this, we quantified how long it takes for the representations from a given time point to reach their peak negative dimensionality. We found that, relative to each training time point, negative dimensionality begins to emerge just after stimulus offset and reaches its peak just before the onset of the next stimulus (Fig. S3), suggesting that the oscillations may be induced by the presentation rate. In contrast, negative correlations are absent from the MEG results. Instead, in the MEG data, the dimensionality scores for cross-time comparisons dissipate as the temporal difference increases, suggesting that the representations gradually decay when there are no intervening stimuli.

We also performed cross-time comparisons of the feature decoding models to see how they compare with the temporal patterns observed for dimensionality (Fig. 3e-f). We focused on the average decoding accuracy of the behavior-derived features. As in the analyses in Figure 2c-d, we fit linear decoding models for each time point in the training data. However, instead of computing decoding scores using only the corresponding time points in the test data, we also computed decoding scores for every cross-time comparison, thus revealing the extent to which a decoder for a given time point generalizes to other times. The temporal patterns in these cross-time decoding results are similar to those in the dimensionality analyses in Figure 3c-d, with a band of strong positive

JNeurosci Accepted Manuscript

decoding scores along the diagonal in both the EEG and MEG data. We also see oscillations in the off-diagonals in the EEG data, fluctuating from positive to negative decoding scores. Thus, these findings demonstrate that the dynamics of feature decoding track with the overall dynamics of the multidimensional neural representations.

Together, these cross-time analyses show that in both the EEG and MEG data, the content of the representations dynamically varies while the dimensionality is nonetheless sustained near its peak level for hundreds of milliseconds after stimulus onset. They further suggest that the structure of the representational dynamics is sensitive to the parameters of the stimulus-presentation timing, with faster and more consistent presentation rates potentially inducing strong oscillations in the stimulus-related responses.

3.4 Leading behavioral and DNN models do not fully account for neural dimensionality

We next sought to characterize the number of dimensions that are shared between the neural representations and the behavioral and DNN features. Our goal was to understand whether existing behavioral and DNN features fully account for the observed neural dimensions or whether the neural representations contain additional stimulus-related information that is not yet captured by these models. To accomplish this, we performed a cross-decomposition analysis, which transforms both the neural responses and features to a shared orthogonal basis where their covariance is maximized (Gauthaman et al., 2025; Han and Bonner, 2026). We used the same training and testing framework as the preceding analyses, with transformations learned on training data and applied to held-out test data. We computed a summary statistic of the shared dimensionality between the neural and feature representations by counting the number of shared dimensions whose reliability was above threshold in the test data. This is analogous to the neural dimensionality metric in Figure 2a-b.

The results are displayed in Figure 4. As expected, the temporal trends are broadly consistent with those observed for decoding accuracy (Fig. 2). We see an earlier peak for the dimensionality of

J Neurosci Accepted Manuscript

low-level DNN features and later peaks for both high-level DNN features and behavioral features. Intriguingly, although the number of features in each set is generally larger than the number of neural channels (except when comparing behavioral embeddings and MEG data), we nonetheless find a substantial gap between the dimensionality of the neural representations themselves and their shared dimensionality with each feature set (Fig. 4). This holds true even when we combine the DNN and behavioral features in a single analysis, and it is consistent across all DNN layers examined (Fig. S4).

In sum, these findings show that leading computational and behavioral models do not yet account for the full complexity of stimulus representations that can be reliably detected in human EEG and MEG data.

3.5 Unexplained variance in neural representations contains behaviorally relevant information

We wondered whether the remaining unexplained variance in neural responses is perceptually relevant. Alternatively, it is possible that all the behaviorally relevant information might be contained within the subspace of the representations that is explained by the existing behavioral embeddings, which were derived from a large-scale dataset of human similarity judgments (Hebart et al., 2020; Hebart et al., 2023a). As a case study, we focused on the EEG responses at 200 ms after stimulus onset, which is around the first peak of behavioral feature decoding. We examined the residual variance in the EEG responses to the test stimuli after regressing out the variance explained by the behavioral embeddings.

We began by visualizing these residual representations in a two-dimensional embedding using uniform manifold approximation and projection (UMAP) (McInnes et al., 2018). As expected, the residual representations do not exhibit readily interpretable clusters (Fig. 5a). This makes sense, since we have already regressed out the most salient factors driving behavioral similarity judgments. However, the remaining representations may nonetheless reflect perceptually relevant information,

JNeurosci Accepted Manuscript

even if it is not easily interpretable.

To test this possibility, we performed a behavioral experiment using the residual neural responses at a sample time point (200 ms). We recruited online participants to perform an image-matching experiment, where the task was to chose which of two clusters a target image should belong to (Fig. 5b). The correct cluster includes five "neighboring" images based on Pearson distance (i.e., one minus correlation of the residual channel responses). The foil cluster includes five random images that are distant from the target. Participants' accuracy was significantly above chance ($M = 0.62$, $SD = 0.06$, $t(49) = 14.63$, $p < 1e-4$, Cohen's d = 2.07; Fig. 5c). We also replicated these findings using the neural representations from another time point (100 ms) to confirm that they are not contingent on the specific time point selected ($M = 0.57$, $SD = 0.05$, $t(49) = 9.34$, $p < 1e-4$, Cohen's d = 1.32; Fig. 5c).

We next asked if the residual neural responses contained only low-level image properties, or whether they also encode high-level information. To address this question, we performed follow-up analyses in which we used the residual responses at sample time points (100, 200, and 300 ms) to decode early- and later-layer DNN features, using the same OpenCLIP ResNet50 model from our previous analyses. We performed these decoding analyses on the residualized responses for both the EEG and MEG data. In both datasets, we find significant decoding of both early and later DNN features at all time points tested (Fig. S5). Furthermore, we find that the nature of the residuals changes over time: when comparing an early time point (i.e., 100 ms) to later time points, we observe relatively stronger decoding of early vs. late DNN features. Thus, these DNN decoding results show that the residual EEG and MEG responses after regressing out previously identified behavioral embeddings, reflect a combination of both low- and high-level information that dynamically changes over time.

Together these findings demonstrate that EEG responses to images encode rich behaviorally relevant information that extends beyond the most salient factors captured by conventional behavioral assessments of perceived similarity. It is worth noting, however, that this behavioral evidence is limited in scope. It was obtained from the EEG data at two time points, and although accuracy

JNeurosci Accepted Manuscript

was reliably above chance across participants, it was modest in absolute terms, as expected for a discrimination that required matching images on residual similarity after the dominant factors driving similarity judgments had been removed.

4 Discussion

We characterized the multidimensional dynamics of human EEG and MEG responses to natural object images using cross-validated matrix decomposition. Our findings reveal that the dimensionality of stimulus-related variance rapidly increases after stimulus onset and is sustained at its peak level for several hundred milliseconds. These neural dimensions are only partially explained by existing behavioral and computational models. Follow-up behavior experiments show that the remaining unexplained variance is behaviorally relevant, with participants reliably matching images based on their neural similarity. These results challenge prevailing assumptions about the inherent low-dimensional limitations of EEG and MEG recordings and reveal rich representational structure beyond what these common models capture.

Our findings have implications for understanding the insights that can be gained from EEG and MEG data. Studies typically focus on leading components or predefined features in these recordings, potentially underestimating their representational richness (Bankson et al., 2018; Carlson et al., 2013; Hansen et al., 2019; Cichy et al., 2014; Grootswagers et al., 2017). Our results show that the stimulus-related information in EEG and MEG data extends well beyond the leading PCs. This suggests a remarkably rich encoding of visual information, consistent with recent work demonstrating complex multidimensional structure in visual cortex representations measured with fMRI (Gauthaman et al., 2025; Han and Bonner, 2026). Moreover, temporally resolved recordings offer unique insights that fMRI studies cannot provide: we can track how the structure of the representations evolves millisecond by millisecond, revealing rapid multidimensional changes as visual processing unfolds. This temporal perspective complements the spatial richness of fMRI, demonstrating that both spatial and temporal resolution contribute distinct and valuable information about

JNeurosci Accepted Manuscript

the dimensionality of neural codes.

Notably, the present analyses characterize dimensionality at individual time points. Future work could seek to reveal even richer multidimensional structure in EEG and MEG data by considering all time points together, as representations may utilize distinct dimensions at different moments throughout visual processing. However, simply concatenating responses across time and applying our decomposition analyses would conflate variance due to stimulus identity with variance due to temporal dynamics alone, making it difficult to identify which dimensions are actually stimulus-related. Alternately, one could concatenate all time points and channels and decompose these spatiotemporal response vectors, but the resulting latent dimensions would mix spatial and temporal information, making it difficult to examine temporal dynamics. Thus, while our work characterizes the multidimensional structure at each time point and the temporal evolution of this structure after stimulus onset, future studies could benefit from the development of new methods that capture the full range of orthogonal dimensions elicited in temporally resolved neural responses.

Our findings also reveal that there is a substantial gap between neural dimensionality and the information captured by conventional representational models (Hebart et al., 2020; Radford et al., 2021; Ilharco et al., 2021). This gap persisted even when combining behavioral embeddings designed to capture human similarity judgments with different layer-wise embeddings of neural networks spanning early to late processing stages. This unexplained variance in EEG and MEG data suggests that current approaches do not yet account for the full complexity of neural representations. The behavioral relevance of this residual neural variance further demonstrates that unexplained dimensions contain meaningful perceptual structure. Participants could reliably match images based on residual neural patterns, showing that the human visual system encodes visual information along dimensions that are not yet fully captured by current models. The behavioral embedding and DNN we tested are widely used models of human visual representation, and given that the brain-aligned representations of DNNs tend to be highly universal (Chen and Bonner, 2025), we interpret this gap as a genuine limitation of existing representational accounts rather than an artifact of model choice.

JNeurosci Accepted Manuscript

Nonetheless, our evidence is limited in scope in two respects. First, the behavioral test used the EEG data at two time points, and the matching accuracies, while reliable across participants, are modest in absolute terms. Second, we have not performed an exhaustive search of existing models, and thus, it remains an open question whether a broad test might reveal a class of model that closes the gap.

Additional factors beyond model quality could contribute to this gap. First, the behavioral embeddings were learned from a different group of participants and, thus, cannot capture potential individual differences in visual representations, which could be distributed across many latent dimensions of neural activity (Han and Bonner, 2026). Second, though we have focused on models of image representation, it is possible that some aspects of the remaining variance could reflect non-sensory factors. Importantly, because our dimensionality estimates depend on correlations across independent stimulus repetitions, anything that varies randomly across repetitions, such as attention lapses, could not add to the dimensionality gap. However, non-sensory factors that are nonetheless reliably stimulus-related could contribute, such as category-specific eye-movement patterns (Caron and Ester, 2026). Thus, an exciting direction for future work is to uncover what key factors underlie the remaining neural variance and to what extent these factors are sensory in nature or might alternatively reflect non-sensory but stimulus-related effects.

Our analyses used orthogonal decomposition methods to characterize the full spectrum of latent dimensions encoding stimulus related information in EEG and MEG data. Previous work has shown that representations from such time-resolved neural data can be mapped to those measured with spatially resolved data, like fMRI, allowing researchers to characterize how stimuli are represented over both time and space (Cichy et al., 2014). However, these previous methods relied on the use of RSA, which is largely driven by the leading PCs of neural responses and may overlook other meaningful but low-variance dimensions (Gauthaman et al., 2025; Han and Bonner, 2026). In contrast, our cross-decomposition method is sensitive to the full spectrum of latent dimensions in neural representations. In future work, this cross-decomposition method could be adapted to

JNeurosci Accepted Manuscript

identify shared dimensions across recording modalities, thus potentially revealing a more detailed picture of the spatiotemporal structure underlying dynamic human brain responses.

Our findings show that, depending on experimental conditions, image representations can exhibit more complex temporal structure than current theories emphasize. Specifically, previous work has largely focused on the distinct temporal profiles observed for different visual features (Isik et al., 2014; Carlson et al., 2013; Cichy et al., 2014; Cichy et al., 2016; Coggan et al., 2016; Grootswagers et al., 2017; Teichmann et al., 2026). We replicate some of these previous findings by showing earlier decoding of low-level features and later decoding of high-level features. However, our findings in the EEG data additionally reveal an unexpected oscillatory pattern in the stimulus encoding, whereby the entire spectrum of representational dimensions oscillates. This means that the stimulus information encoded at one time point is also observed at later time points but with the opposite polarity. We suspect that this transient oscillatory pattern could be induced by the rapid and regular temporal structure of stimulus presentation in the EEG experiment. However, even if this explanation is correct, it raises important questions for future work. First, it remains unknown what functional relevance these oscillations might have, if any. One possibility is that they are related to oscillatory mechanisms of visual attention (Fiebelkorn and Kastner, 2019). Second, it highlights the possibility that conventional representational analyses, such as feature decoding, could easily overlook major representational transformations if they are applied to each time point independently or on temporally averaged data.

Our work reveals that time-resolved neural responses to natural images contain rich multidimensional structure, which in the EEG data approaches the ceiling set by the number of recording channels. This structure cannot be fully explained by current representational models, whether derived from similarity judgments or deep neural networks, yet it contains behaviorally relevant information. By revealing rich structure in modalities often considered noisy or limited, our results open new avenues for investigating the neural basis of vision. Together, these findings highlight the importance of going beyond the leading dimensions of neural activity and understanding dynamic

JNeurosci Accepted Manuscript

human brain responses in their full complexity.

JNeurosci Accepted Manuscript

5 Data Availability

The EEG dataset (THINGS-EEG2) is publicly available at (Gifford et al., 2022a; Gifford et al., 2022b). The MEG dataset (THINGS-MEG) is publicly available at (Hebart et al., 2023a; Hebart et al., 2023b). Anonymized behavioral data collected in this study are available at (Chen et al., 2026a).

6 Code Availability

The code for reproducing all analyses and figures in this manuscript is available at (Chen et al., 2026b).

7 Declaration of Interests

The authors declare no competing financial interests.

JNeurosci Accepted Manuscript

References

Allen EJ, St-Yves G, Wu Y, Breedlove JL, Prince JS, Dowdle LT, Nau M, Caron B, Pestilli F, Charest I, Hutchinson JB, Naselaris T, Kay K (2022) A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience* 25:116–126.

Bankson BB, Hebart MN, Groen II, Baker CI (2018) The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage* 178:172–182.

Barrett LF, Miller EK (2025) Categorization is "baked" into the brain. *PsyArXiv* .

Carlson T, Tovar DA, Alink A, Kriegeskorte N (2013) Representational dynamics of object vision: the first 1000 ms. *Journal of vision* 13:1–1.

Caron A, Ester EF (2026) Human gaze behaviors track abstract stimulus categories. *Journal of Cognitive Neuroscience* 38:1513–1526.

Chang N, Pyles JA, Marcus A, Gupta A, Tarr MJ, Aminoff EM (2019) BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific data* 6:49.

Chen Z, Bonner MF (2025) Universal dimensions of visual representation. *Science Advances* 11:eadw7697.

Chen Z, Isik L, Bonner MF (2026a) Behavioral data for: Multidimensional dynamics of object representations in the human visual system Zenodo https://doi.org/10.5281/zenodo.21937065.

Chen Z, Isik L, Bonner MF (2026b) Code for: Multidimensional dynamics of object representations in the human visual system GitHub https://github.com/zche377/temporal_dimensionality.

Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports* 6:27755.

JNeurosci Accepted Manuscript

Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. *Nature neuroscience* 17:455–462.

Coggan DD, Baker DH, Andrews TJ (2016) The role of visual and semantic properties in the emergence of category-specific patterns of neural response in the human brain. *ENeuro* 3.

Connolly AC, Guntupalli JS, Gors J, Hanke M, Halchenko YO, Wu YC, Abdi H, Haxby JV (2012) The representation of biological classes in the human brain. *Journal of Neuroscience* 32:2608–2618.

Doerig A, Kietzmann TC, Allen E, Wu Y, Naselaris T, Kay K, Charest I (2025) High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence* 7:1220–1234.

Fiebelkorn IC, Kastner S (2019) A rhythmic theory of attention. *Trends in cognitive sciences* 23:87–101.

Fusi S, Miller EK, Rigotti M (2016) Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology* 37:66–74.

Gauthaman RM, Ménard B, Bonner MF (2025) Universal scale-free representations in human visual cortex. *PLOS Computational Biology* 21:e1013714.

Gifford AT, Dwivedi K, Roig G, Cichy RM (2022a) A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage* 264:119754.

Gifford AT, Dwivedi K, Roig G, Cichy RM (2022b) THINGS-EEG2 dataset OSF https://osf.io/anp5v.

Greene MR, Hansen BC (2018) Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLoS computational biology* 14:e1006327.

Greene MR, Hansen BC (2020) Disentangling the independent contributions of visual and conceptual features to the spatiotemporal dynamics of scene categorization. *Journal of Neuro-*

JNeurosci Accepted Manuscript

science 40:5283–5299.

Grootswagers T, Robinson AK, Shatek SM, Carlson TA (2019) Untangling featural and conceptual object representations. *NeuroImage* 202:116083.

Grootswagers T, Wardle SG, Carlson TA (2017) Decoding dynamic brain patterns from evoked responses: a tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of cognitive neuroscience* 29:677–697.

Han C, Bonner MF (2026) High-dimensional structure underlying individual differences in naturalistic visual experience. *Current Biology* 36:723–733.

Hansen BC, Field DJ, Greene MR, Olson C, Miskovic V (2019) Towards a state-space geometry of neural responses to natural scenes: A steady-state approach. *NeuroImage* 201:116027.

Hebart MN, Contier O, Teichmann L, Rockter AH, Zheng CY, Kidder A, Corriveau A, Vaziri-Pashkam M, Baker CI (2023a) THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife* 12:e82580.

Hebart MN, Contier O, Teichmann L, Rockter AH, Zheng CY, Kidder A, Corriveau A, Vaziri-Pashkam M, Baker CI (2023b) THINGS-MEG dataset OpenNeuro https://openneuro.org/datasets/ds004212/versions/3.0.0.

Hebart MN, Dickter AH, Kidder A, Kwok WY, Corriveau A, Van Wicklin C, Baker CI (2019) THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one* 14:e0223792.

Hebart MN, Zheng CY, Pereira F, Baker CI (2020) Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour* 4:1173–1185.

Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neu-*

JNeurosci Accepted Manuscript

ron 76:1210–1224.

Ilharco G, Wortsman M, Wightman R, Gordon C, Carlini N, Taori R, Dave A, Shankar V, Namkoong H, Miller J, Hajishirzi H, Farhadi A, Schmidt L (2021) OpenCLIP Zenodo https://doi.org/10.5281/zenodo.5143773.

Isik L, Meyers EM, Leibo JZ, Poggio T (2014) The dynamics of invariant object recognition in the human visual system. *Journal of neurophysiology* 111:91–102.

Khosla M, Murty NAR, Kanwisher N (2022) A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology* 32:4159–4171.

Lehky SR, Kiani R, Esteky H, Tanaka K (2014) Dimensionality of object representations in monkey inferotemporal cortex. *Neural computation* 26:2135–2162.

Lehky SR, Tanaka K (2016) Neural representation for object recognition in inferotemporal cortex. *Current opinion in neurobiology* 37:23–35.

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. *Journal of neuroscience methods* 164:177–190.

McInnes L, Healy J, Saul N, Grossberger L (2018) UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software* 3:861.

McMahon E, Im EJ, Bonner MF, Isik L (2025) A spatiotemporal hierarchy for social interaction perception in the lateral visual stream. *PsyArXiv* .

Michel CM, Brunet D (2019) EEG source imaging: a practical review of the analysis steps. *Frontiers in neurology* 10:325.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning transferable visual models from natural language supervision In *International conference on machine learning*, pp. 8748–8763. PmLR.

JNeurosci Accepted Manuscript

Slapik MB, Shouval HZ (2026) Simulated complex cells contribute to object recognition through representational untangling. *Neural Computation* 38:145–164.

Stringer C, Pachitariu M, Steinmetz N, Carandini M, Harris KD (2019) High-dimensional geometry of population responses in visual cortex. *Nature* 571:361–365.

Teichmann L, Hebart MN, Baker CI (2026) Dynamic representation of multidimensional object properties in the human brain. *Journal of Neuroscience* .

Thorat S, Proklova D, Peelen MV (2019) The nature of the animacy organization in human ventral temporal cortex. *elife* 8:e47142.

Wang AY, Kay K, Naselaris T, Tarr MJ, Wehbe L (2023) Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence* 5:1415–1426.

JNeurosci Accepted Manuscript

Figure legends

Figure 1. Overview of datasets and metrics evaluating neural representations and visual features of interest. **(a)** Left: Both EEG (Gifford et al., 2022a) and MEG (Hebart et al., 2023a) datasets are analyzed in this work and visual features of interest are vector representations related to these natural object images, with examples shown here. The stimuli are from the THINGS database, which contains images from 1,854 object categories (Hebart et al., 2019). Right: The two neural datasets have different experimental designs: the EEG dataset presented images for 100 milliseconds (ms) and had 100 ms of fixation in between, while the MEG dataset presented images for 500 ms and had 1 s of fixation in between. **(b)** Left: Sparse nonnegative embeddings (SPoSE) were learned to explain human perceived similarity judgments in an odd-one-out behavioral experiment. Right: Features were extracted from layers of a ResNet50 model, trained on the YFCC15M dataset with the CLIP objective (Radford et al., 2021; Ilharco et al., 2021). Two layers were selected to capture low- and high-level visual features, and these DNN representations were reduced using sparse random projections. **(c)** We assessed the dimensionality of image representations at each individual time point. Further, to compare neural data and model features, we computed both shared-space dimensionality—a symmetric method using cross-decomposition (Gauthaman et al., 2025; Han and Bonner, 2026)—and decoding scores—a traditional asymmetric method using ridge regression. subj, subjects; ch, channels.

Figure 2. Temporal dynamics of neural dimensionality and feature decoding. **(a, b)** These plots show the dimensionality of stimulus-related variance in the EEG and MEG datasets at each time point following stimulus onset. **(c, d)** These plots show the average decoding scores for behavioral embeddings and neural network features at each time point. The vertical lines show the stimulus-presentation timing of the two datasets, with solid lines indicating stimulus onsets and dashed lines indicating offsets. The error bars for dimensionality represent the standard error across subjects and those for decoding scores represent the standard error of feature-averaged means across subjects. The two datasets have unequal sample sizes, as indicated in the panel titles (EEG,

JNeurosci Accepted Manuscript

n = 10; MEG, n = 4 subjects), so the MEG standard errors are estimated from fewer participants. The EEG and MEG panels cover different time ranges but are drawn at the same width, so equal horizontal distances do not correspond to equal durations. For both datasets, the dimensionality increases sharply after stimulus onset and is sustained from 100-300 ms in the EEG data and 150-400 ms in the MEG data before gradually declining. All three sets of features can be decoded from the neural responses, and the decoding scores are generally highest during time points of high neural dimensionality. Horizontal lines under the decoding time courses indicate time windows of significant decoding ($p < 0.05$, cluster-based permutation test). subj, subjects; ch, channels.

Figure 3. Temporal generalization of dimensionality and feature decoding. Throughout this figure, time axes are expressed relative to stimulus onset, where train time is the time point at which the PCs or decoding models were learned and test time is the time point at which they were evaluated. **(a), (b)** These plots illustrate how the PCs from three example time points evolve over time. At each time point relative to stimulus onset, these plots show the number of dimensions that have significant positive correlations with PCs learned from the selected example time points. The results show that in both the EEG and MEG data, representational PCs from earlier, intermediate, and later time points exhibit distinct temporal trajectories. They also show that the EEG data exhibit more rapid representational changes, consistent with the faster stimulus-presentation timing in the EEG experiment. The error bars represent the standard error across subjects; as in Figure 2, the two datasets have unequal sample sizes (EEG, n = 10; MEG, n = 4 subjects), as indicated in the panel titles. The EEG and MEG panels cover different time ranges but are drawn at the same width, so equal horizontal distances do not correspond to equal durations. **(c), (d)** These matrices show the number of dimensions shared across time points, with visualizations for both positively and negatively correlated dimensions. Positive dimensionality counts the number of dimensions along which two time points have significant positive correlations, while negative dimensionality counts the number of significant negative correlations. Each cell displays the greater value of the positive and negative dimensionality metrics. Both datasets exhibit a band of strong positive dimensionality along the diagonal, and the EEG data additionally exhibit prominent oscillations

JNeurosci Accepted Manuscript

from positive to negative dimensionality. **(e), (f)** These matrices show the average decoding scores for behavioral features, and they include all cross-time comparisons. Cells marked with dashes have non-significant decoding scores ($p < 0.05$, cluster-based permutation test). The trends in these matrices are broadly consistent with the dimensionality trends. Both datasets exhibit a band of strong positive decoding along the diagonal, and the EEG data additionally exhibit prominent oscillations from positive to negative decoding scores. subj, subjects; ch, channels.

Figure 4. Temporal dynamics of shared dimensions between neural representations and visual features. These plots show the shared dimensionality between neural representations at each time point and the behavioral and DNN features. The dimensionality metric reflects the number of significant shared dimensions identified from cross-decomposition analyses comparing neural representations with behavioral and DNN features. The results labeled "everything" are for cross-decomposition analyses in which the behavioral and DNN features were combined. These plots also include the dimensionality of the neural representations themselves, which are also shown in Figure 2. The vertical lines show the experimental setup of the two datasets, with solid lines indicating stimulus onsets and dashed lines indicating offsets. The error bars represent the standard error across subjects. The two datasets have unequal sample sizes (EEG, $n = 10$; MEG, $n = 4$ subjects), as indicated in the panel titles. The EEG and MEG panels cover different time ranges but are drawn at the same width, so equal horizontal distances do not correspond to equal durations. These analyses reveal a substantial gap between the dimensionality of the neural representations themselves and the shared dimensionality with the behavioral and DNN features. subj, subjects; ch, channels.

Figure 5. Behavioral relevance of residual variance in neural data after partitioning out embeddings learned from large-scale human similarity judgments. (a) This plot shows a UMAP embedding of the residual variance in EEG responses at 200 ms after partialling out the behavioral features learned from a large-scale dataset of human similarity judgments (Hebart et al., 2020; Hebart et al., 2023a). As expected, the most salient interpretable structure in the data has been

JNeurosci Accepted Manuscript

partialled out, but the remaining unexplained variance may nonetheless reflect perceptually relevant information. A behavioral experiment was designed to test this possibility. **(b)** In the behavioral experiment, participants ($N = 50$) chose which group the target images should belong to. The correct group includes five neighboring images with the closest Pearson distance to target image in the residual representations, while the foil cluster includes five random images that are distant from the target image. **(c)** These violin plots show the distribution of accuracy scores across participants at each of the two time points tested ($N = 50$ in each experiment). The 200 ms time point corresponds to the residual representations visualized in (a), and the 100 ms time point is a replication. Participant accuracy is significantly above chance at both time points, showing that the residual variance in the neural data reflects perceptually relevant information and that this result is not contingent on the specific time point selected. The dashed line marks chance performance. **** $p < 1e-4$

JNeurosci Accepted Manuscript

JNeurosci Accepted Manuscript

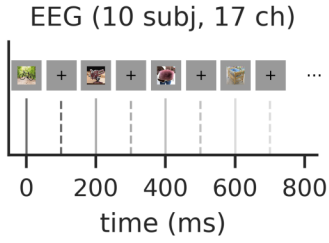
a

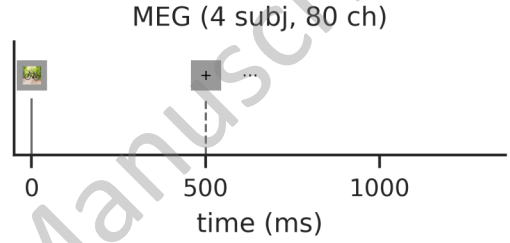
stimuli

dynamic neural data



1,854 object categories
EEG: 16,740 unique images
MEG: 20,048 unique images





b

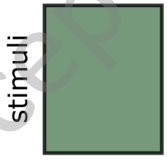
features of interest

① behavior embeddings

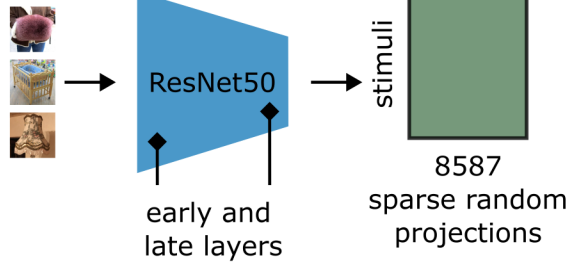
"Which is the odd one out?"



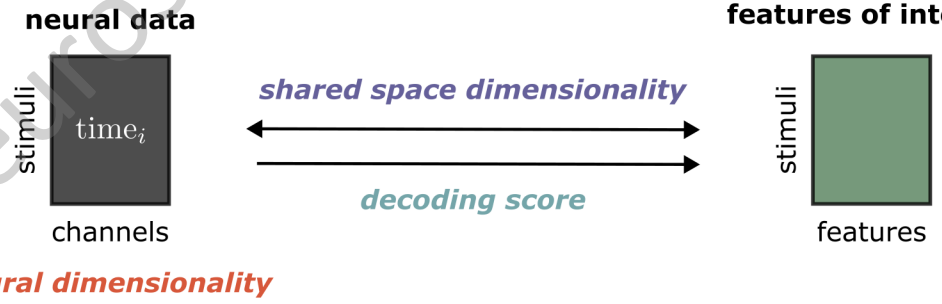
SPoSE

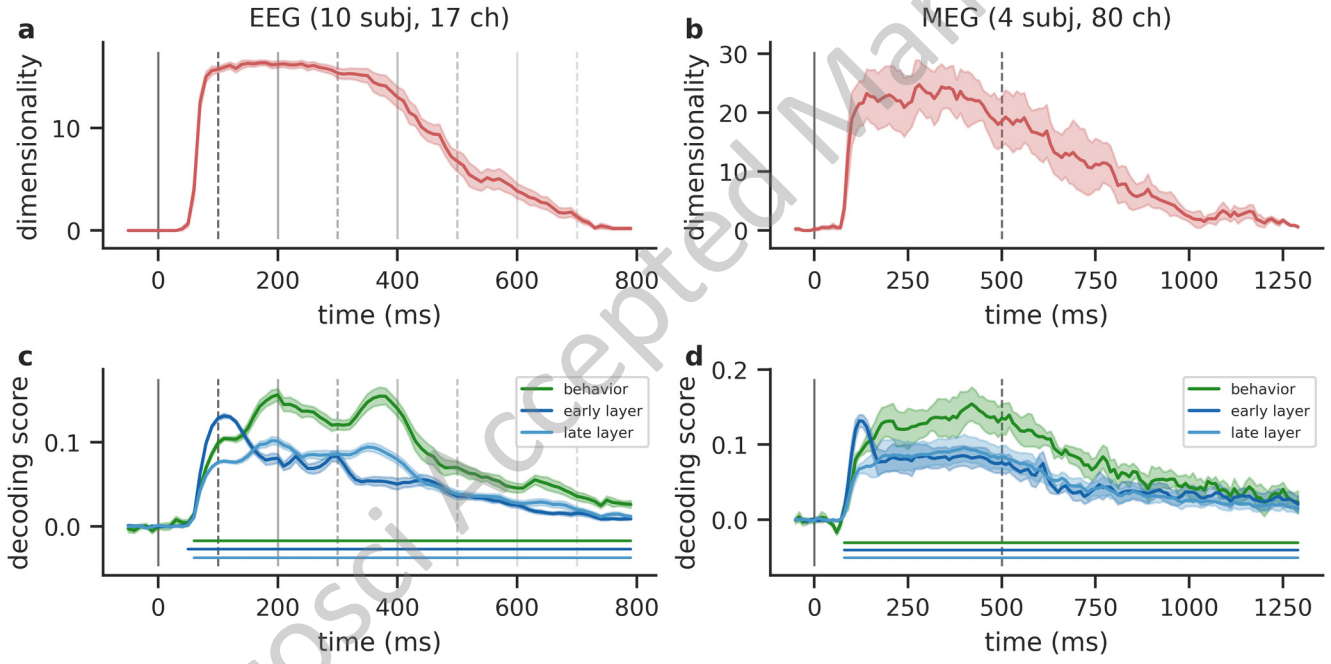


② neural network feature maps

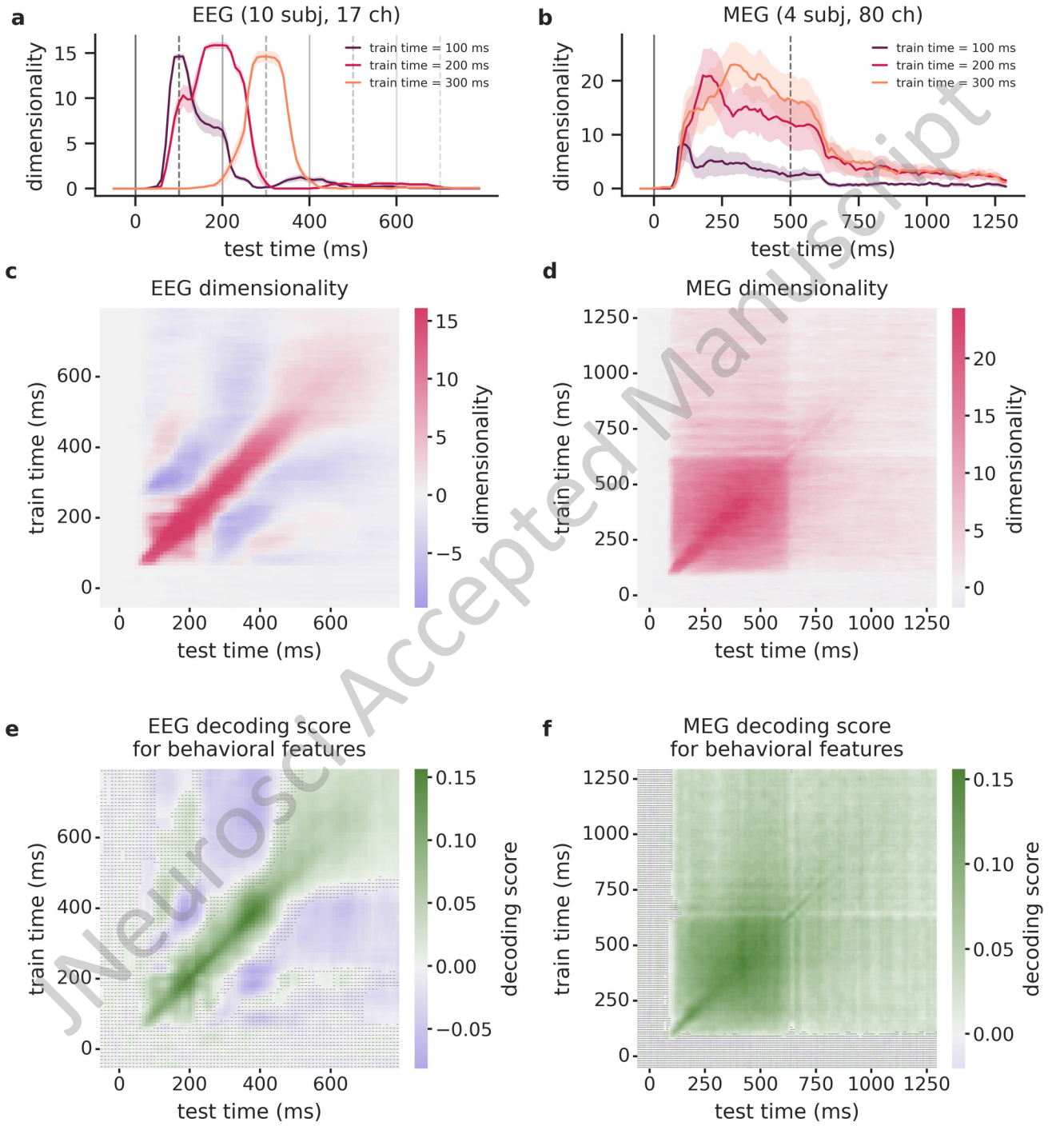


c

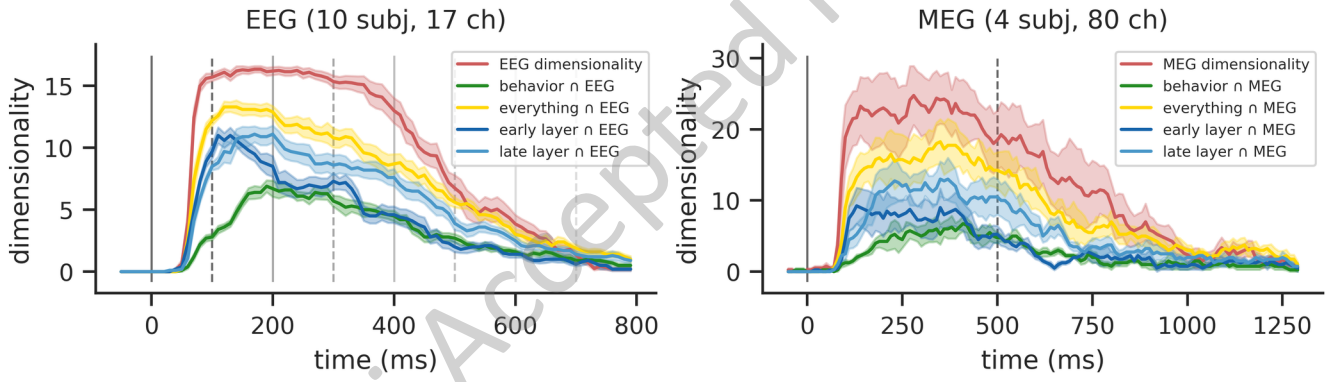




Neurosci Accepted Manuscript

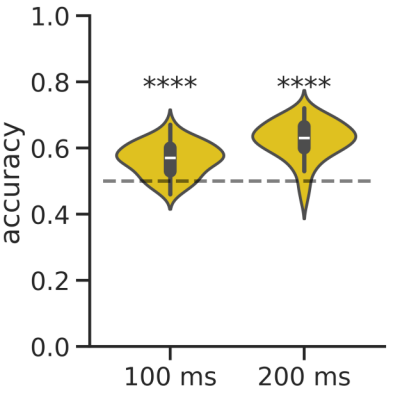


J Neurosci Accepted Manuscript


EEG (10 subj, 17 ch)
dimensionality
time (ms)
EEG dimensionality
behavior ∩ EEG
everything ∩ EEG
early layer ∩ EEG
late layer ∩ EEG
MEG (4 subj, 80 ch)
dimensionality
time (ms)
MEG dimensionality
behavior ∩ MEG
everything ∩ MEG
early layer ∩ MEG
late layer ∩ MEG







Neuroscience Peer Reviewed Manuscript