Behavioral relevance of high-dimensional neural representations

Chihye Han (chan21@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218

Raj Magesh Gauthaman (rgautha1@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218

Michael F. Bonner (mfbonner@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218

Abstract

A common approach to understanding the organizing principles of neural representations has been to emphasize high-variance dimensions that correspond to interpretable features. Here, we investigate whether behavioral relevance is restricted to these interpretable dimensions or spans the entire spectrum of neural representations. Using fMRI data from the Natural Scenes Dataset, we tested whether humans could perceive coherent structure in image clusters formed along principal components of ventral visual stream responses, where explained variance decreases by orders of magnitude across principal-component ranks. In this initial study examining the first two decades of neural dimensions, we found that behavioral relevance extends throughout the entire range tested. These findings suggest that behaviorally relevant information in neural representations extends beyond the interpretable, high-variance dimensions emphasized in standard approaches and that comprehensive models of neural coding should account for the full range of dimensions.

Introduction

How should we understand neural representations in human visual cortex? Existing work has gained success in identifying a set of interpretable neural dimensions that map onto perceptually distinct or behaviorally relevant stimulus categories and properties. These approaches characterize neural populations in terms of interpretable properties, such as object and action labels (Huth et al., 2012), social features (Tarhan & Konkle, 2020), reachspace classes (Josephs, Hebart, & Konkle, 2023), and material descriptors (Schmidt et al., 2025). Similarly, Hebart et al. derived interpretable object dimensions from human similarity judgments and used them to chart the tuning profiles of visual cortex (Hebart et al., 2020; Contier, Baker, & Hebart, 2024). These works focus on the explanatory value of interpretable dimensions as a way to model neural responses, as they reveal the key factors underlying our ability to make sense of our visual world, to structure our external environment, and to act on it (Contier et al., 2024).

However, recent work suggests that neural representations are high-dimensional and extend far beyond the few dimensions that can be interpreted visually or semantically (Gauthaman, Ménard, & Bonner, 2024; Han & Bonner, 2025). One key finding from this line of work is that conventional variance-weighted approaches capture only a limited portion of the representational space, effectively ignoring many lower-variance dimensions that may be functionally significant (Gauthaman et al., 2024; Haxby et al., 2011). According to this view, neural representations are best understood statistically in their full dimensionality.

How can we reconcile the former view that emphasizes the functional significance of interpretable dimensions and the latter view that provides support for the inherent high-dimensionality of neural representations? One possibility is that even though neural representations are high-dimensional, only the low-dimensional subset contains perceptually structured and thus behaviorally relevant information. Alternatively, behavioral relevance may span the entire spectrum of neural dimensions, even in dimensions with low interpretability. We address this question by examining whether humans can perceive coherent structure in image clusters formed along neural dimensions. If only a few interpretable components carry behaviorally relevant information, then humans should only reliably detect structure that is encoded by these high-variance, semantically interpretable dimensions. If behavioral relevance extends throughout the spectrum, humans may identify coherence even in dimensions that lack obvious interpretability.

Methods

To test if people can identify latent structure encoded by neural dimensions, we sought to create stimuli that are representative of each neural dimension. To do so, we used fMRI data from one subject in the Natural Scenes Dataset (Allen et al., 2022), leveraging high-resolution neural responses to tens of thousands of natural scenes. We focused on responses in the midventral visual stream and averaged across three scan repetitions.

Stimuli sampling We applied principal component analysis (PCA) to neural responses and identified images with maximal (highest or lowest) values along each resulting latent dimension. We constructed target stimuli sets with the images located at each dimension's "pole" and selected foil stimuli by randomly sampling images that do not form clusters along the same dimension (Fig. 1**A**). To assess whether the structure of each latent dimension has significant behavioral relevance, we designed a simple perceptual judgment task, where participants were presented with both target and foil groups of images arranged in grids and asked to identify which group contained more similar images (Fig. 1**B**). We verified that our stimulus selection procedure created coherent target clusters that were specific to each dimension by computing the ratio between average pairwise within-cluster distances among target stimuli sampled at poles and within-cluster distances among random foils. The ratio was lowest for the dimensions from which the clusters were derived, confirming that cluster coherence was unique to each component (Fig. 1**C**). Fig. 1**D** shows example sets of stimuli. The target image set for Dimension 6 consists of human and animal faces, akin to a canonical "face dimension." In contrast, the target images for Dimension 90 as well as random foils from these dimensions include a mixture of categories with no apparent organizing principle.

Coherence judgment experiment We measured the behavioral relevance of neural dimensions by assessing whether participants could discriminate coherent image clusters formed along those dimensions. For each dimension, we quantified behavioral relevance as the mean accuracy with which participants identified target over foil groups.

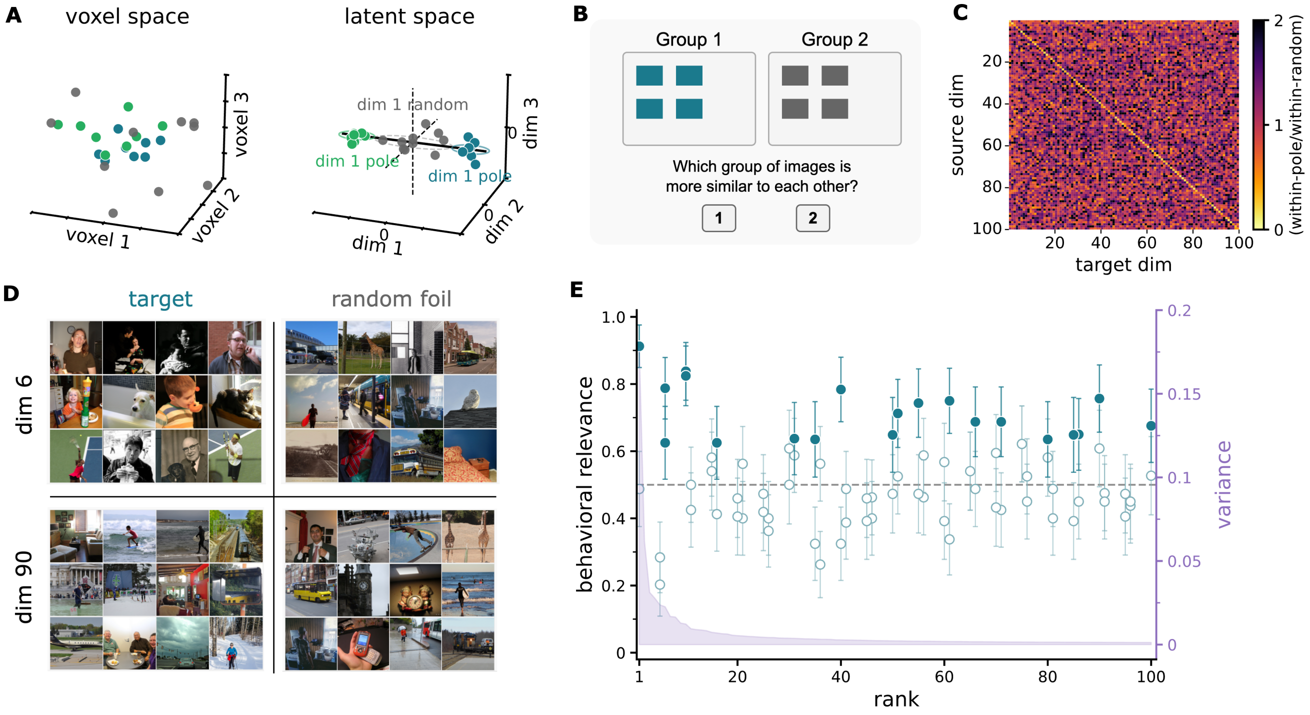

Figure 1: **A. Structure in latent dimensions.** Neural responses to natural images in voxel space (left) are projected into latent space (right). Clusters at dimensional poles represent the structure encoded along each dimension. **B. Coherence judgment task.** The target group (teal) consists of images sampled from a dimensional pole; the foil group (grey) consists of randomly sampled images from the same dimension. Participants indicate which group has more similar images. **C. Validation of coherence structure.** Heatmap values show average distances of within-pole images divided by average distances of within-random images. Lower values indicate greater coherence in pole clusters. The diagonal pattern shows that coherence is specific to the source dimensions. **D. Example stimuli.** Target and foil clusters from dimensions 6 (top) and 90 (bottom). Both dimensions showed reliable behavioral relevance in our experiment. **E. Behavioral relevance of dimensions.** Behavioral relevance (teal; filled = statistically significant, $p < 0.05$) and principal-component explained variance in neural data (purple). Behaviorally relevant dimensions are distributed across the entire range tested. Error bars: 95% CI. Dotted line: chance level.

Participants were recruited via Prolific for online participation. We conducted two experiments testing non-overlapping sets of dimensions from the first 100 dimensions: the first experiment included twenty dimensions sampled evenly from dimensions 1 and 96 and the second included twenty dimensions sampled evenly from dimensions 5 to 100. Participants who missed more than one catch trial of simple shape similarity were excluded (1 and 4 participants for Experiments 1 and 2, respectively), leaving 80 and 74 subjects for final analysis.

Results & Discussion

Out of 40 dimensions tested (ranks 1-100), 18 showed statistically significant above-chance behavioral relevance (one-way t-test; $p < 0.05$, FDR-corrected). Importantly, these behaviorally relevant dimensions were found across the spectrum, even though the majority of variance in neural responses was explained by the first few dimensions (Fig. 1**E**). While explained variance decreases rapidly across dimensions, behavioral relevance does not follow the same decay.

Our findings suggest that interpretability—the ability to assign semantic labels or visually distinctive features to dimensions—is not a necessary condition for behavioral relevance. Low-variance dimensions that are not subject to semantic readout contain meaningful structure and potentially contribute to downstream behavior. The striking dissociation between variance and behavioral relevance suggests that the neural code utilizes information available across the full range of dimensions, not just a subset of interpretable high-variance dimensions.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., . . . others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Contier, O., Baker, C. I., & Hebart, M. N. (2024). Distributed representations of behaviour-derived object dimensions in the human visual system. *Nature Human Behaviour*, *8*(11), 2179–2193.

Gauthaman, R. M., Ménard, B., & Bonner, M. F. (2024). *Universal scale-free representations in human visual cortex.* Retrieved from `https://arxiv.org/abs/2409.06843`

Han, C., & Bonner, M. F. (2025). *High-dimensional structure underlying individual differences in naturalistic visual experience.* Retrieved from `https://arxiv.org/abs/2505.12653`

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., . . . Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*(2), 404–416.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Josephs, E. L., Hebart, M. N., & Konkle, T. (2023). Dimensions underlying human understanding of the reachable world. *Cognition*, *234*, 105368.

Schmidt, F., Hebart, M. N., Schmid, A. C., & Fleming, R. W. (2025). Core dimensions of human material perception. *Proceedings of the National Academy of Sciences*, *122*(10), e2417202122.

Tarhan, L., & Konkle, T. (2020). Sociality and interaction envelope organize visual action representations. *Nature Communications*, *11*(1), 3002.