# Is visual cortex really "language-aligned"? Perspectives from Model-to-Brain Comparisons in Human and Monkeys on the Natural Scenes Dataset

**Colin Conwell (conwell@g.harvard.edu)**
Johns Hopkins University
Department of Cognitive Science

**Emalie MacMahon**
Johns Hopkins University
Department of Cognitive Science

**Kasper Vinken**
Harvard Medical School
Department of Neurobiology

**Saloni Sharma**
Harvard Medical School
Department of Neurobiology

**Akshay Jagadeesh**
Harvard Medical School
Department of Neurobiology

**Jacob S. Prince**
Harvard University
Department of Psychology

**George A. Alvarez**
Harvard University
Department of Psychology

**Talia Konkle**
Harvard University
Department of Psychology

**Leyla Isik**
Johns Hopkins University
Department of Cognitive Science

**Margaret Livingstone**
Harvard Medical School
Department of Neurobiology

## Abstract

**Recent progress in multimodal AI and "language-aligned" visual representation learning has re-ignited debates about the role of language in shaping the human visual system. In particular, the emergent ability of "language-aligned" vision models (e.g. CLIP) – and even pure language models (e.g. BERT) – to predict image-evoked brain activity has led some to suggest that human visual cortex itself may be "language-aligned" in comparable ways.**

**But what would we make of this claim if the same procedures worked in the modeling of visual activity in a species that has no language? Here, we deploy controlled comparisons of pure-vision, pure-language, and multimodal vision-language models in prediction of human (N=4) and rhesus macaque (N=6, 5:IT, 1:V1) ventral visual activity evoked in response to the same set of 1000 captioned natural images (the "NSD1000"). Preliminary results reveal markedly similar patterns in aggregate model predictivity of early and late ventral visual cortex across both species. Together, these results suggest that language predictivity of the human visual system is not necessarily due to "language-alignment" *per se*, but rather to the statistical structure of the visual world as reflected in language.**

**Introduction** The idea that language shapes how we 'see' the world has been the focus of an almost century-long debate in cognitive (neuro)science (7; 19; 15; 11; 6), and has evolved through many forms over that time. A recent evolution of this idea has taken the form of a debate about the extent to which high-level human visual cortex is 'language-aligned' – or, in other words, the extent to which lingustic or linguistically-learned structure is evident in visual brain responses (13; 16). The resurgence of this debate is predicated in large part on two seminal findings: first, the finding that 'language-aligned' machine vision models (e.g. CLIP) are some of the most predictive models to date of image-evoked activity in the visual brain (18); and second, the finding that even pure-language models (e.g. BERT) are capable of predicting image-evoked activity brain activity by way of image captions alone (4; 17).

Here, we apply a logical razor to this argument in the form of assessing whether these two key findings hold in the brain of a species that does not speak language. We call this the 'monkey razor': If the ability of 'language-aligned' vision models or pure-language models to predict image-evoked brain activity is indeed evidence of language having (re-)shaped visual representation, we should not find similar trends in monkeys.

**Methods** Our approach is to use encoding models fit to the feature spaces of a diverse set of pure-vision (VMs), pure-language (LMs), and multimodal (language-aligned) vision (VLMs) models to predict image-evoked brain activity in the ventral stream of 4 humans and 6 rhesus macaques shown the same set of 1000 natural images from the Natural Scenes Dataset (NSD) (1). Our encoding procedure follows an established protocol for large-scale model comparison (3), and includes a nested cross-validation regime that decontaminates the selection of the most brain-like layer *within* each of our candidate models (assessed on a 'training set' of 500 images) from the comparison of brain-likeness *between* models (assessed on the held-out 'test set' of 500 images). The predictivity of the encoding models is assessed with the raw Pearson correlation ($r$) between model-predicted and actual brain activity. (The calculation of comparable noise ceilings across species for use in a measure of 'explainable variance explained' (2; 14) is an area of ongoing investigation).

The brain-likeness of the pure-vision and language-aligned vision models is assessed on the images themselves. The brain-likeness of the pure-language models is assessed using an average of the embeddings for the first 5 captions associated with each image as part of the Microsoft COCO metadata (10) (from which NSD images are curated). The encoding models for the human (fMRI) brain activity are fit to reliability-selected voxels ($NCSNR > 0.2$) in a broad mask of early visual cortex (EVC, N=15326 voxels) and occcipitemporal cortex (OTC, N=29840 voxels), with both anatomical and functional criteria as the basis of inclusion. The encoding models for the monkey (electrophysiology) brain activity are fit to multi-unit responses (i.e. average firing rates in a 150ms window) from arrays placed either in macaque V1 (N=34 units) or inferotemporal (IT) cortex (N=394 units). Note that we use the following convention for reporting statistics: statistic [lower, upper] 95% (bootstrapped) confidence interval.

## Results

**Vision versus Language in Human Ventral Stream** (Top Row of Figure 1) Commensurate with previous findings (8; 5), we find that pure-language model embeddings over image captions are sufficient to predict high-level human ventral visual activity almost as accurately as pure-vision models, with mean voxel-wise OTC encoding scores of $r$=.332 [.298, .375], .365 [.336, .395], and .365 [.336, .393] for pure-language, pure-vision, and language-aligned vision models, respectively. Conversely, language models perform far worse than pure-vision and language-aligned vision models in prediction of early ventral stream activity, with mean voxel-wise EVC encoding scores of $r$=.178 [.153, .219], .335 [.302, .356], and , respectively.

**Vision versus Language in Macaque Ventral Stream** (Bottom Row of Figure 1) Applying the same encoding procedures with the same probe stimuli to prediction of brain activity in macaque visual cortex, we find (as in humans) that pure-language models are remarkably accurate in predicting high-level ventral visual activity, with mean IT encoding scores of $r$=.343 [.266, .427]. Also as in humans, we find that pure-language models perform poorly in prediction of early visual cortex ($r = 0.107$ for the single V1 subject), and that there is no substantial difference between pure-vision models and language-aligned vision models, with mean IT encoding scores of $r = .441$ [.36, .523] and .415 [.343, .488]. There is, however, a slightly more pronounced difference between the pure-language and pure-vision models in macaque IT (.343 versus

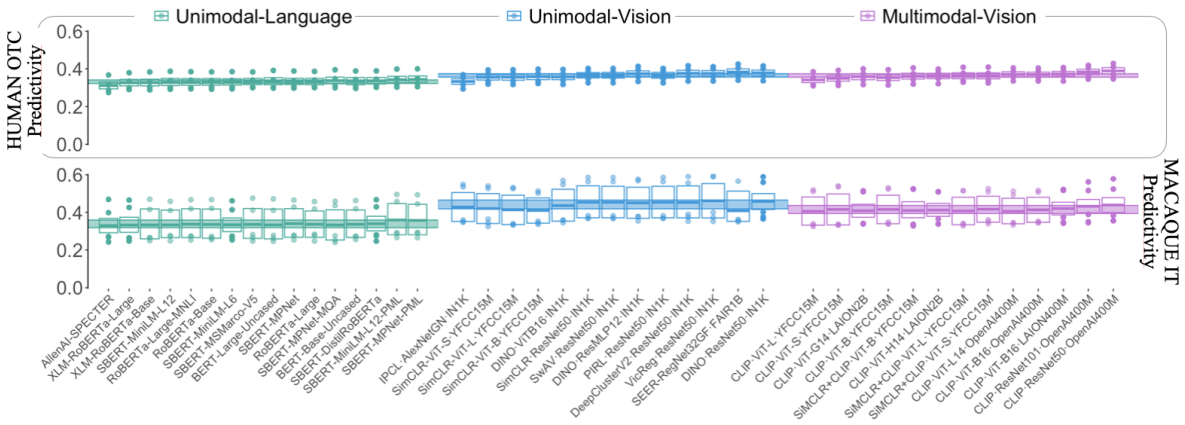


Figure 1: Encoding accuracies from the most brain-like layer of a series of (unimodal) pure-language, pure-vision, and (multimodal) vision-language models in prediction of both human occipitotemporal cortex (OTC) and macaque inferotemporal cortex (IT). Indiviudal points are the accuracies for individual subjects (human or macaque). The vertical striped boxes are the means $\pm$95%CIs (across subjects) per model. The horizontal, semitranslucent rectangles extending over these striped boxes are the means $\pm$95% BCIs (across models) per model type (modality).

.441) compared to human OTC (.332 versus .365).

**Cross-Encoding Analysis of Interspecies Difference** This divergence in the pattern of results between the two species may be due to multiple factors, including the species-specific recording modalities and preprocessing stesp (electrophysiology versus functional imaging). The question of most relevance, here, though, is whether the difference is attributable to language (or at least, language as encoded in the pure-language models). To assess this directly, we performed a modified cross-encoding analysis in which we used the macaque IT data to predict the human OTC data, using the exact same encoding procedure we originally applied to the model features. We then used either the most OTC-predictive pure-language embeddings or the most OTC-predictive pure-vision embeddings to predict *the residuals* of this (interspecies) cross-encoding model to see how much either modality could account for the unshared structure across species. The logic of this analysis is that if the difference between humans and monkeys is a difference attributable to language, then the pure-language embeddings should more accurately predict the variance that remains in human brains once we've accounted for the structure that is shared with monkey brains.

Applying this analysis, we find first that macaque IT data is reasonably accurate in predicting human OTC data, with an average voxelwise-encoding score of $r$=.25 [0.19, 0.30] across subjects, confirming prior reports that a sizable portion of the representational structure in OTC is shared with macaque IT (12; 9). Subsequently, we find that pure-language models are in fact worse on average than pure-vision models in predicting the residuals of the monkey-human cross-encoding, with Pearson correlations between predicted and actual residuals of $r$=.095 [.069, .12] for pure-language models and $r$=.187 [.151, .223] for pure-vision models. This analysis suggests (at minimum) that the unshared structure between the species is *not* uniquely defined by structure learned through language.

**Summary** In this preliminary work, we show that the ability of 'language-aligned' vision models and pure-language models to predict image-evoked brain activity in human high-level visual cortex is likely not evidence of language having reshaped vision. We find similar trends using these models to predict visual brain responses in a species that has no language, demonstrating as well that those differences which do exist between humans and monkeys are not directly attributable to the structures of language alone. Such is the nature of the 'monkey razor': If, *caeteris paribus*, an experimental effect holds in both humans and monkeys, that effect cannot be attributable to the structure, function, deployment, or learning of language *per se*. Thus, the more likely explanation here is that the representational overlap between pure-language models and the high-level primate visual brain reflects an organization of the world that is intuitive to us long before we learn to speak a language – a structure learnable in large part through the hierarchical encoding of natural image statistics. Language (as learned by language models) may approximate the representational endpoints of this process, but only to the extent that these statistics are reflected in the language we use to describe the world around us (a world the language models themselves cannot actually 'see').

Further work is needed to make sense of the lingering difference (however small) between language model predictivity of human OTC and macaque IT. One major factor that merits further scrutiny here is the translation between different neural recording modalities: fMRI signals, for example, may include later visual components (including feedback) not evident in the electrophysiological signal. In future work, we hope to assess the tradeoff between pure-language and pure-vision model encoding over time – an analysis that could unveil an even greater degree of similarity between humans and macaques than the initial similarity we've shown here. Perhaps more importantly, we could also aspire to collect or curate visual brain data in both species that pushes the limits of representation learnable through image statistics alone – and extends more explicitly into the kinds of conceptual territories where the structures of language are most indispensable for understanding.

## 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. (Publisher: Nature Publishing Group US New York) doi: 10.1038/s41593-021-00962-x

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2023). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*. doi: 10.1101/2022.03.28.485868

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*.

Doerig, A., Sommers, R., Seeliger, K., Richards, B., Ismael, J., Lindsay, G., ... others (2022). The neuroconnectionist research programme. *arXiv preprint arXiv:2209.03718*.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and brain sciences*, *39*, e229.

Hoijer, H. E. (1954). Language in culture; conference on the interrelations of language and other aspects of culture.

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.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).

Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*, *24*(4), 279–284.

Orban, G. A., Van Essen, D., & Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends in cognitive sciences*, *8*(7), 315–324.

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, *24*(11), 1628–1636. (Publisher: Nature Publishing Group) doi: 10.1038/s41593-021-00921-6

Pospisil, D. A., & Bair, W. (2021). The unbiased estimation of the fraction of variance explained by a model. *PLoS computational biology*, *17*(8), e1009212.

Rosch, E. (2015). Linguistic relativity. In *Human communication* (pp. 95–121). Routledge.

Seydell-Greenwald, A., Wang, X., Newport, E. L., Bi, Y., & Striem-Amit, E. (2023). Spoken language processing activates the primary visual cortex. *PLoS One*, *18*(8), e0289671.

Tang, J., Du, M., Vo, V., Lal, V., & Huth, A. (2024). Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems*, *36*.

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & 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*(12), 1415–1426.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, *104*(19), 7780–7785.