CONTROLLED ASSESSMENT OF CLIP-STYLE LANGUAGE-ALIGNED VISION MODELS IN PREDICTION OF BRAIN & BEHAVIORAL DATA

Colin Conwell *, **Jacob S. Prince, Chris Hamblin & George A. Alvarez**
Harvard University, Department of Psychology
{conwell,prince,hamblin,alvarez}@{g,fas,fas,wjh}.harvard.edu

ABSTRACT

One of the core algorithmic forces driving the development of modern foundation models is the use of contrastive language alignment to facilitate more robust visual representation learning. The clear benefits conferred by CLIP-style multimodal objective functions in computer vision have generated a frenzy of interest in the application of these models to a long-debated question in cognitive neuroscience: to what extent does language shape perceptual representation in the human mind? In this work, we explore this question in two distinct domains: the prediction of brain activity in the human ventral visual system (as measured by high-resolution fMRI), and the prediction of visually evoked affect in human image assessment (as measured by self-report). In both of these cases, we leverage popular open-source foundation models (e.g. OpenAI's CLIP) in conjunction with empirically controlled alternatives (e.g. Meta AI's SLIP models) to better isolate the effects of language alignment while holding architecture and dataset constant. These controlled experiments offer mixed evidence regarding the influence of language on perceptual representation: specifically, when architecture and dataset are held constant, we find no evidence that language-alignment improves the brain predictivity of vision models, but we do find strong evidence that it increases predictivity of behavioral image assessments. We offer these examples as a case study in the urgency of injecting greater empirical control into the development and evaluation of foundation models, whose emergent properties may be attributable to a variety of sources that only systematic model comparison can fully disentangle.

1 INTRODUCTION

To what extent does human language shape visual representation and function? Some theories propose that high-level visual cortical representations can be explicitly shaped by linguistic and semantic concepts (13; 22; 6; 35; 31; 21). The top-down influence of language on vision is especially clear during category learning in development (30; 43). Alternate theories place heavier emphasis on the structure of natural image statistics as representational constraints, and relatively less weight on the role of top-down factors (3; 19; 28; 24; 29). Indeed, recent work suggests that object category information can be learned through purely domain-general learning of the covariance structure of natural images, and that these representations can explain substantial variance in neural datasets (23).

The advent of foundation models now offers a valuable opportunity to gain further traction on this theoretical debate. Prominent successes in accounting for neural responses in high-level visual cortex have occurred using models trained almost exclusively on unimodal visual datasets, such as ImageNet (see 25; 40 for review). Foundation models trained on massive multimodal datasets now offer a direct avenue to compare how sensory and linguistic constraints affect representation learning at scale, and to assess their relative importance in accounting for rich profiles of brain and behavioral responses. A recent study of the popular OpenAI Contrastive Language-Image Pre-training (CLIP) model revealed that it explained a greater degree of unique variance in large-scale fMRI recordings than did a canonical vision-only model (ImageNet-trained ResNet-50, 42). However, it remains unclear whether CLIP's superior predictive capacity arises from the semantic information contained in its training image captions, or alternatively, from the sheer magnitude or quality of its training dataset relative to ImageNet. Given that training dataset size and diversity may indeed be the primary factors underlying models' neural and behavioral predictivity (9; 11), the goal of the present study is to further understand the specific impacts of language-alignment on both brain and

*Corresponding author: conwell@g.harvard.edu

behavioral prediction, while showcasing model comparison methods that more tightly control the many other kinds of inductive biases that may impact these kinds of predictions more generally, including model architecture and training dataset.

2 METHODS

Brain and Behavioral Predictions

The brain data we use in this analysis consists of a subset of the large-scale 7T fMRI Natural Scenes Dataset (NSD) (2): specifically, activity from 44806 voxels in the ventral visual stream of 4 subjects responding to 1000 natural images from the COCO image set. We focus our analysis on three anatomical sectors of interest: early visual cortex (defined as voxels falling within dorsal and ventral V1, V2, V3, and hV4 masks); occipitotemporal cortex (defined as voxels that cover the ventral and lateral object-responsive cortex, including category-selective regions); and, the visual word-form area (VWFA), a specific subset of OTC that shows highly selective responses to written characters. (For details, see Figure 2A and Appendix A.2).

The behavioral data we use are ratings of visually evoked affect (arousal, valence, and beauty) from the OASIS dataset (26; 7): a set of 900 images rated on a scale of 1 to 7 for each of the 3 affect ratings. Each image is rated by 100-110 human raters, responding to prompts such as: "How positive or negative does this image make you feel?" We take the group-average ratings per image per affect as the main target of our analysis.

To predict the brain and behavioral data using our candidate deep neural network models, we use cross-validated reguralized linear regressions computed across dimensionality-reduced feature spaces extracted from each layer of each model (10; 11). We use the cross-validated max predictions across layers as the overall score for each model in a given dataset. (For details, see Appendix A.4).

Candidate Neural Network Models

In predicting both the brain and behavioral data, we test a large battery of deep neural network models (N = 145) from a variety of sources. These models were largely hand-selected to vary meaningfully across three core dimensions: input (training data), computational architecture, and learning objective. The main model contrast we assess in this dataset is the contrast between unimodal (purely visual) models, and multimodal (language-aligned vision) models. (We take unimodal models to mean models trained purely with visual self-supervision; we do not include category-supervised models trained with language-adjacent one-hot encoding vectors for category labels).

The majority of the contrasts between the unimodal (vision-only) and multimodal (language-aligned vision) models in our model set are *uncontrolled*, in that they vary across more than one dimension (training data, architecture, or learning objective) at a time. Our main *controlled* empirical contrasts are between the models of the SLIP-family, which include a purely visual self-supervised model (SLIP-SimCLR) that shares an architecture (ViT-[S,B,L]) and training set (YFCC15M) with two multimodal variants (SLIP-CLIP, and SLIP-Combo, the latter of which combines visual contrastive learning with language alignment). The SLIP models are crucial to this analysis in that they allow us to isolate the effects of language alignment, holding architecture and dataset constant. (See Appendix A.3 for details).

Crucially, we note that because both of our target human datasets in this work are image-based (in that they consist solely of images, and a matrix of human brain or behavioral responses evoked by each image), we extract representations only from the visual backbones of those models that contain multimodal components (e.g. CLIP, which has both a visual and textual backbone). In other words, our probe stimuli are not themselves multimodal; the multimodality in this analysis comes only in the form of learned representations in our candidate, pretrained models.

After computing feature regression scores for all our models on both of our target datasets, we use rank statistics and direct model-to-model comparisons to probe for the influence of language across unimodal (vision-only) and multimodal (language-aligned vision) model candidates.

Brain and Behavioral Modeling: Experimental Logic

Worth emphasizing in more detail here is the overarching logic of our experiments. To assess the extent to which language shapes visual representation in humans, we seek machine vision models

that learn similar representations *with or without* objectives that involve language – ideally in such a way that the only difference between these models is precisely the influence of language. (This is what we mean by unimodal, vision-only models versus multimodal, language-aligned vision-models, and by 'controlled' comparisons that ensure the only difference between these models is, in fact, some form of language-alignment). Our unimodal (vision-only) models (e.g. SimCLR, SwaV, BarlowTwins) tend to be contrastive-learning models whose visual representations are shaped exclusively by the learning of visual selectivity and invariance across augmentations of individual image instances. No language is involved in this learning. Our multimodal (language-aligned) vision models, on the other hand, are models that learn by simultaneously embedding both visual and textual inputs, and aligning the embedded representations of both. Language, in this case, is usually the primary shaper of the visual representations these models eventually learn. Our interpretation, then, as to whether a given target of prediction (i.e. brain or behavioral) is influenced by language, comes from the differential predictivity scores of the models that learn with or without it.

3 RESULTS

These analyses address two primary questions: (1) In *uncontrolled* comparisons of unimodal (vision-only) and multimodal (language-aligned vision) models, which may differ along multiple dimensions (architecture, dataset, and task), what models best predict human brain and behavioral data? (2) *Controlling for architecture and dataset*, does multimodality in the form of language-alignment affect the model's ability to predict our target brain and behavioral data? The key data points we leverage to answer these questions are: (1) the performance of the top-ranked unimodal model relative to the top-ranked multimodal model (whatever their underlying architectures and training data); and (2) the performance of the unimodal (SimCLR) SLIP variant relative to its multimodal counterparts (CLIP and Combo). Unless otherwise noted, we use the following convention in the reporting of scores: arithmetic mean [lower, upper 95% bootstrapped confidence interval] across subjects.

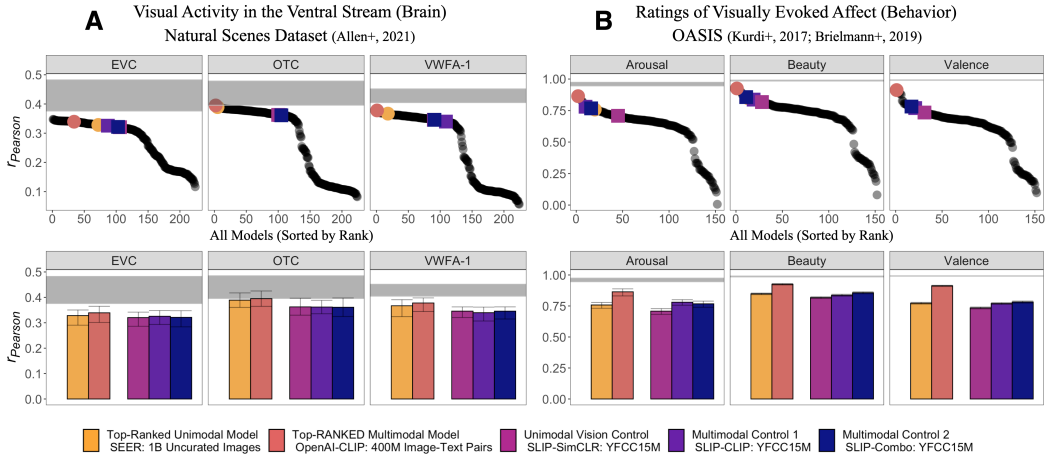

Figure 1: A summary of our main results: In the top panels of both these plots, you can see the relative ranking of our target models (colored points) with respect to the full set of models (black points) we surveyed. In the bottom panel, we zoom in on the scores of our target models. The translucent gray bars in both plots are the noise ceilings we calculate for each dataset (NCSNR (2) for the fMRI data, and Spearman-Brown splithalf reliability for the image ratings). The error bars are bootstrapped confidence intervals over subjects. **A** In the visual system, the differences between the top-ranking unimodal model and the top-ranking multimodal model (yellow and salmon bars to the left), as well as the differences across the 3 architecture-and-dataset-controlled SLIP variants (pink, purple, and blue to the right) are negligible in all cortical areas we assess. **B** In ratings of visually evoked affect, the top-ranked multimodal model far outperforms the top-ranked unimodal model. The statistically significant superiority of the multimodal SLIP variants SLIP-CLIP and SLIP-Combo confirms this difference is very likely attribute to the language alignment inherent to their learning objectives.

Predicting Neural Responses in the Human Visual System (Brain)

We first turn to our brain data, where we find that even our uncontrolled model comparisons show no evidence of language-aligned models outperforming unimodal (vision-only) models in predictions of human visual cortical responses. In occipitotemporal cortex, the highest ranked model (out of our target 145 models) was a category-supervised ResMLP architecture trained on Imagenet21k, with a mean accuracy of $r_{Pearson}$ = 0.398 [0.369, 0.428]. The 2nd highest ranked model was the ResNet50 variant of OpenAI's CLIP model, with a mean voxelwise-encoding accuracy of $r_{Pearson}$ = 0.395 [0.365, 0.425]. Ranked 5th out of 145 models in its prediction of OTC voxel activity is a RegNet-64Gf variant of FaceBook's SEER algorithm (17) – a large-scale, purely-visual, contrastive learning model, with mean voxelwise-encoding accuracy of 0.389 [0.361, 0.418]. Though the difference between the two models here ($\Delta r_{Pearson}$ = -0.0058 [-0.0092, -0.003]) is technically significant in that it replicates across all subjects, the effect is so miniscule as to make the difference meaningless.

The controlled comparison between the highest performing unimodal and multimodal SLIP (ViT-B) variants makes it even more apparent that language alignment confers no advantage in predicting ventral stream responses. Consider the difference between SLIP-SiMCLR and SLIP-Combo (the sole difference in this case being the addition of the language alignment objective, since both use visual augmentations as part of their learning): Though the ranks of these models are relatively low (at rank 84 and 87, respectively), the scores of each model are decent (SLIP-SimCLR $r_{Pearson}$ = 0.363 [0.330, 0.397]; SLIP-Combo $r_{Pearson}$ = 0.361 [0.325, 0.398], and the difference between ($\delta r_{Pearson}$ = 0.000115 [-0.00842, 0.00736]) them is insignificant (bootstrapped $p = .39$).

What these comparisons make clear is that multimodality in the form language alignment DOES NOT seem to confer meaningful benefits when predicting activity in high-level visual cortex. This story is much the same in early visual cortex (where even the most predictive OpenAI-CLIP model ranks 31st / 154 models), but also, perhaps more surprisingly, in the visual word form area (a region of cortex that preferentially responds to written characters, and that only develops this selectivity after we learn to read.) Though OpenAI's ResNet50-CLIP is the highest ranked model in this region of cortex (mean voxelwise-encoding accuracy of $r_{Pearson}$ = 0.378 [0.344, 0.398]), RegNet-64Gf-SEER is again only minimally behind ($r_{Pearson}$ = 0.367 [0.324, 0.391]), and the controlled comparison of SLIP-SimCLR in this region is almost numerically identical to SLIP-Combo: 0.346 [0.321, 0.362] vs. 0.346 [0.315, 0.363].

In short, the representational structure of human visual cortex (as measured by the NSD) is captured *equally well* by unimodal (vision-only) and multimodal (language-aligned vision) models. (See Figure 1A for details).

Predicting Human Visually Evoked Affect (Behavior)

This equivalence disappears as soon as we move from the characterization of visual representation in the brain to the outputs of visual representation at the level of behavior – at least, to the outputs of visual representation combined with whatever conceptual, linguistic, or contextual variables dictate our affective responses to visual stimuli.

In the uncontrolled comparisons across the 145 models we test in this analysis (varying along multiple characteristics), OpenAI's CLIP-ViT-L/14 is by a large margin the most predictive model of all 3 affect ratings available to us in the OASIS dataset, ranking 1st in predictions of beauty and valence, and 2nd in predictions of arousal (in 1st place for arousal is OpenAI's CLIP-ViT-B/16 variant). Averaging across the 3 measures of affect, in this case, CLIP-ViT-L/14 scores a remarkable average of $r_{Pearson}$ = 0.907 [0.917, 0.912]. The highest ranking *unimodal* model in predictions of affect is again RegNet-64Gf-SEER, which ranks 20th / 154, and scores a far lower average of $r_{Pearson}$ = 0.785 [0.763, 0.794]. This difference ($\Delta r_{Pearson}$ = 0.108 [0.0732, 0.145]) is significant at $p < 0.0001$ (proportion of 10000 bootstrapped resamples in which the difference is greater than 0, Holm-corrected for multiple comparisons), and constitutes an effect that corresponds to approximately 20% of the explainable variance in this data.

In controlled comparisons between SLIP models, the highest-ranking architecture-matched SLIP models across all 3 predictions of affect are the ViT-B variants. Looking again at the difference in accuracy between the SLIP-SimCLR and SLIP-Combo model (which vary only in their language alignment), we see that the SLIP-SimCLR model (on average) predicts human ratings with an accuracy of $r_{Pearson}$ = 0.782 [0.773, 0.790]. The SLIP-Combo model, on the other hand, predicts those

same ratings with an accuracy of $r_{Pearson}$ = 0.817 [0.810, 0.824]. These findings suggest that there is a small, but reliable advantage for language-aligned models when predicting human affect ratings.

In short, the controlled comparisons across the SLIP models do show an advantage for language-alignment. (See Figure 1B for details). Taken together with the vastly superior OpenAI's CLIP-ViT-L/14 performance, these findings suggest the intriguing possibility that the benefits of language for predicting human visually affect may accumulate with scale. (Only with further comparisons across larger, *controlled* foundation model sets can we confirm, however, that this is indeed the case.)

4 Discussion

Given these results, it is tempting to offer a number of direct interpretations about the impact of language on perceptual representation. That language alignment *doesn't* seem to matter for predicting activity in the ventral visual stream could, for example, suggest that the formation of high-level visual representations is encapsulated from the higher-order cognitive processes of reasoning and inference that language (at least in humans) most significantly dictates (14; 15). That it *does* seem to matter for predicting visually evoked affect is in some sense an easy extension of this same point: By the time otherwise encapsulated, feedforward visual representations are reinterpreted through the lens of 'feeling', and all the physiological change and cognitive abstraction that feeling almost always entails, the involvement of language seems almost a given (4). (Participants in affect labeling experiments (41) such as these are, after all, prompted with written instructions.)

Beyond any one specific interpretation, however, these results underscore the importance of comparing models (like the SLIP models) with *controlled variation* – especially when those comparisons involve large-scale foundation models whose 'emergent properties' may ultimately factor into any number of downstream applications (8; 5), including the kinds of brain and behavioral modeling we've done here. More fully characterizing those properties by empirically disentangling the algorithmic pressures that produce them (architecture, data quality and scale, learning objective) is a necessary step on the road to building foundation models that are more robust, more interpretable, and more intelligent. Only once we've built such models can we more fully leverage them to understand the various aspects of human perception and language we hope they'll one day mimic.

References

[1] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 274–281, 2001.

[2] Emily Jean Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Logan T Dowdle, Bradley Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, et al. A massive 7t fmri dataset to bridge cognitive and computational neuroscience. *bioRxiv*, 2021.

[3] Michael J Arcaro and Margaret S Livingstone. On the relationship between maps and domains in inferotemporal cortex. *Nature Reviews Neuroscience*, 22(9):573–583, 2021.

[4] Lisa Feldman Barrett and Eliza Bliss-Moreau. Affect as a psychological primitive. *Advances in experimental social psychology*, 41:167–218, 2009.

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[6] Stefania Bracci, J Brendan Ritchie, and Hans Op de Beeck. On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia*, 105:153–164, 2017.

[7] Aenne A Brielmann and Denis G Pelli. Intense beauty requires intense pleasure. *Frontiers in psychology*, 10:2420, 2019.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] Colin Conwell, D Graham, and EA Vessel. The perceptual primacy of feeling: Affectless machine vision models robustly predict human visual arousal, valence, and aesthetics, 2021.

[10] Colin Conwell, Daniel Graham, and Edward A Vessel. The perceptual primacy of feeling: Affectless machine vision models robustly predict human visual arousal, valence, and aesthetics, Sep 2021. URL psyarxiv.com/5wg4s.

[11] Colin Conwell, Jacob S Prince, George Alvarez, and Talia Konkle. Large-scale benchmarking of diverse artificial vision models in prediction of 7t human neuroimaging data. *bioRxiv*, pp. 2022–03, 2022.

[12] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[13] Hans P Op de Beeck, Ineke Pillet, and J Brendan Ritchie. Factors determining where category-selective areas emerge in visual cortex. *Trends in cognitive sciences*, 23(9):784–797, 2019.

[14] Chaz Firestone and Brian J Scholl. Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and brain sciences*, 39:e229, 2016.

[15] Jerry A Fodor. *The modularity of mind*. MIT press, 1983.

[16] Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. Vissl. https://github.com/facebookresearch/vissl, 2021.

[17] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022.

[18] Trevor Hastie and Robert Tibshirani. Efficient quadratic regularization for expression arrays. *Biostatistics*, 5(3):329–340, 2004.

[19] Daniel Janini and Talia Konkle. A pokémon-sized window into the human brain. *Nature human behaviour*, 3(6):552–553, 2019.

[20] William B Johnson. Extensions of lipschitz mappings into a hilbert space. *Contemp. Math.*, 26:189–206, 1984.

[21] Frederik S Kamps, Cassandra L Hendrix, Patricia A Brennan, and Daniel D Dilks. Connectivity at the origins of domain specificity in the cortical face and place networks. *Proceedings of the National Academy of Sciences*, 117(11):6163–6169, 2020.

[22] Nancy Kanwisher. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, 107(25):11163–11170, 2010.

[23] Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022.

[24] Talia Konkle and Aude Oliva. A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, 74(6):1114–1124, 2012.

[25] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

[26] Benedek Kurdi, Shayn Lozano, and Mahzarin R Banaji. Introducing the open affective standardized image set (oasis). *Behavior research methods*, 49(2):457–470, 2017.

[27] Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 287–296, 2006.

[28] Margaret S Livingstone, Michael J Arcaro, and Peter F Schade. Cortex is cortex: Ubiquitous principles drive face-domain development. *Trends in cognitive sciences*, 23(1):3, 2019.

[29] Bria Long, Chen-Ping Yu, and Talia Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38):E9015–E9024, 2018.

[30] Gary Lupyan, David H Rakison, and James L McClelland. Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological science*, 18(12):1077–1083, 2007.

[31] Bradford Z Mahon and Alfonso Caramazza. What drives the organization of object knowledge in the brain? *Trends in cognitive sciences*, 15(3):97–103, 2011.

[32] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[35] Marius V Peelen and Paul E Downing. Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia*, 105:177–183, 2017.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[37] Ryan M Rifkin and Ross A Lippert. Notes on regularized least squares. 2007.

[38] Alexander Sax, Bradley Emi, Amir R. Zamir, Leonidas J. Guibas, Silvio Savarese, and Jitendra Malik. Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. 2018.

[39] Alexander Sax, Jeffrey O Zhang, Bradley Emi, Amir Zamir, Silvio Savarese, Leonidas Guibas, and Jitendra Malik. Learning to navigate using mid-level visual priors. *arXiv preprint arXiv:1912.11121*, 2019.

[40] Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual review of vision science*, 5:399–426, 2019.

[41] Jared B Torre and Matthew D Lieberman. Putting feelings into words: Affect labeling as implicit emotion regulation. *Emotion Review*, 10(2):116–124, 2018.

[42] Aria Yuan Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *BioRxiv*, pp. 2022–09, 2022.

[43] Sandra R Waxman and Dana B Markow. Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, 29(3):257–302, 1995.

[44] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[45] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[46] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.

A APPENDIX

A.1 OVERVIEW OF METHODS

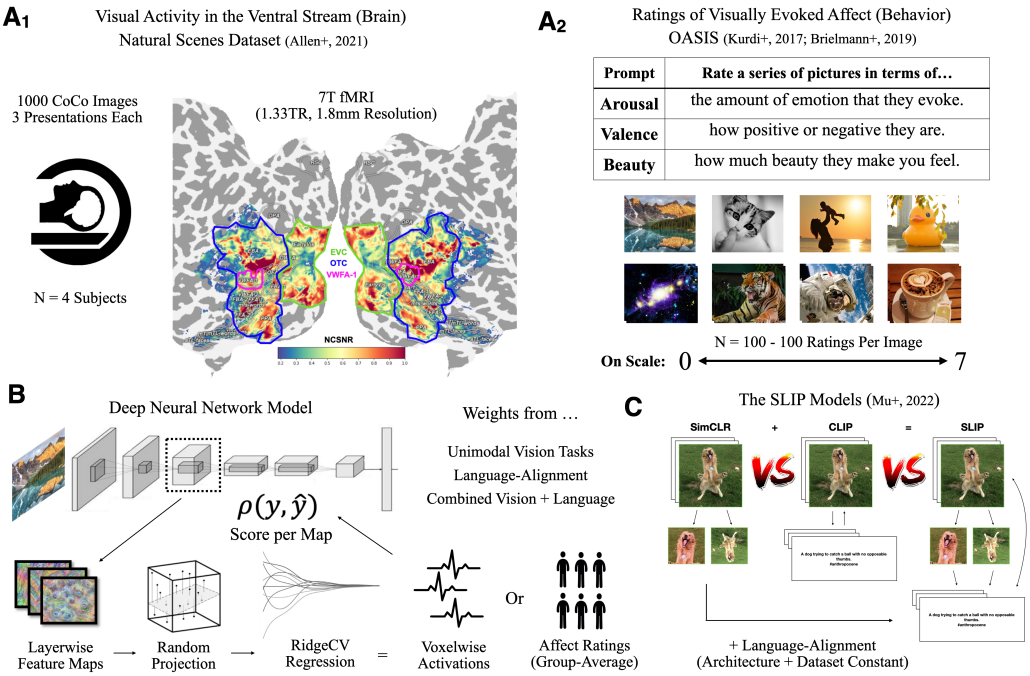

Figure 2: $\mathbf{A}_1$ provides a summary of our target brain data: activity measured in 44806 voxels across the ventral visual cortices of 4 subjects viewing 1000 natural images from the COCO image set. $\mathbf{A}_2$ provides a summary of our target behavioral data: 900 images ranked by 100 to 110 human raters tasked with evaluating each image in terms of its evoked arousal, valence, or beauty. **B** shows a schematic of our pipeline for predicting voxel activity or image ratings from our candidate deep neural network models: extraction of features from each distinct layer, dimensionality reduction of these features using sparse random projection, deployment of these dimensionality-reduced features in a cross-validated regularized regression with the target brain or behavioral data as output. **C** shows a schematic of the SLIP models – our main empirical model contrast. In these models, the *sole and unique* difference between SimCLR (a unimodal visual contrastive learning model) and SLIP (a combined vision-language contrastive learning model) is language alignment: Architecture (ViT) and dataset (YFCC15M) are held constant.

A.2 DETAILS ON THE BRAIN DATASET

The Natural Scenes Dataset (2) contains measurements of 73,000 unique stimuli from the Microsoft Common Objects in Context (COCO) dataset (Lin et al., 2014) at high resolution (7T field strength, 1.33s TR, $1.8mm^3$ voxel size). In this analysis, we focus on the brain responses to 1000 COCO stimuli that overlapped between subjects, and limit analyses to the 4 subjects (subjects 01, 02, 05, 07) for whom all 3 image repetitions are available for the overlapping images. The 3 image repetitions were averaged to yield the final voxel-level response values in response to each stimulus.

Voxel Selection Procedure

Between the 4 NSD subjects analyzed here, there were 44806 total voxels that entered into the neural encoding procedure. For all analyses using NSD data, we analyzed voxels with a noise-ceiling signal-to-noise ratio (NCSNR) > 0.2. For the character-selective ROI, we additionally applied a selectivity threshold, only including voxels with a localizer t-value > 1 (as assessed using independent data). A visualization of these ROIs from a representative subject is shown in Figure 2A, with voxelwise NCSNR values plotted on the cortical surface.

A.3 Details on Model Selection + Sources

Across both our behavioral and brain datasets, we survey a combined total of 173 distinct models (224 when including randomly-initialized variants of key architectures). In this analysis, we focus on a subset of these models (N = 145) that have been tested on *both* the behavioral and brain dataset.

These models are sourced from multiple different repositories, including the Torchvision (PyTorch) model zoo (33); the pytorch-image-models (timm) library (44); the VISSL (self-supervised) model zoo (16); the CLIP collection (36); the Taskonomy (visualpriors) project (46; 38; 39); and the Detectron2 model zoo (45). The first two of these repositories offer pretrained versions of a large number of object recognition models with varying architectures: including (classic and modern) convolutional networks, vision transformers, and MLP-mixers. For each of these 'ImageNet' (object recognition) models, we include one trained and one randomly initialized variant (using whatever initialization scheme the model authors recommend) so as to assess the impact of ImageNet training on brain prediction, and as a sanity check. The self-supervised models are mainly variants on a popular convolutional architecture (ResNet-50), though do include some transformers (e.g. the DINO ViT and XCiT models). The Taskonomy models consist of a core encoder-decoder architecture trained on 24 different common computer vision tasks, ranging from autoencoding to edge detection. These models are engineered in such a way that only the architecture of the decoder varies across task, allowing us to assess (after detaching the encoder) what effect different kinds of training has on brain predictivity, independent of model design.

Our primary goal in this analysis was to contrast unimodal (purely visual) models, and models that learn from vision and language simultaneously. Our unimodal model candidates consisted almost entirely of self-supervised algorithms from the VISSL model zoo (16), including Jigsaw, RotNet, NPID+, PIRL, SimCLR, SwaV, Dino and SEER. Importantly, *none* of these models are trained using linguistic targets, even in the form of the one-hot encoding vectors used to train object recognition models. Our primary multimodal model candidates consist of the 7 ResNet + ViT visual backbones from OpenAI's CLIP repository.

Importantly, almost all of these unimodal and multimodal in this uncontrolled set differ along more than one of our 3 axes (training data, architecture, and learning objective). CLIP, in particular, is trained on a heretofore proprietary dataset of 400 million image-text pairs that none of the unimodal models are trained on. Attributing differences between CLIP and other models to language alignment alone, then, is empirically dubious.

To address this, we use Meta AI's SLIP models (32) – a series of Vision Transformers (Small [ViT-S], Base [ViT-B], & Large [ViT-L]), all trained on the YFCC15M dataset (15 million image-text pairs), but on only 1 of 3 tasks: pure SimCLR-style self-supervision; pure CLIP-style language alignment; or the eponymous SLIP – a combination of self-supervision and language alignment.

A.4 Feature Regression Procedure

In this appendix, we provide further detail on the methods we use to predict brain and behavioral data from the feature spaces of our candidate deep neural network models.

A.4.1 Regressing Voxel-wise Activations (Brain Data)

Feature Extraction We aim to understand which models represent the voxel responses to images explicitly. To do this, we fit linear regressions from features in the latent layers of neural networks to voxel responses. Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{1000}$ denote the 1000 COCO images, and $\mathbf{Y} = \{y_i\}_{i=1}^{1000}$ the corresponding scalar responses of a given voxel. For a given model, we first extract feature vectors for each of its latent layers. For layers that return a tensor per input of more than one dimension, such as convolutional layers, we flatten the tensor into a vector. We treat the application of an activation function as a distinct layer with its own feature vector. In the case of transformers, we consider the key - query - value feature vectors independently as well. Let's denote the array of feature vectors a given layer returns in response to n images as $\mathbf{H} \in \mathbb{R}^{n \times m}$, where m is the dimensions of each feature vector.

Sparse Random Projections For some deep-net layers m is very large, and as such performing ridge regression directly on $\mathbf{H}$ is prohibitively expensive, with at best linear complexity with m, $\mathcal{O}(n^2 m)$ (18). This is especially problematic considering we are regressing to many voxels from many latent layers across many models. Fortunately it follows from the Johnson-Lindenstrauss

lemma (20; 12) that $\mathbf{H}$ can be projected down to a low-dimensional embedding $\mathbf{P} \in \mathbb{R}^{n \times p}$ that preserves pair-wise distances of points in $\mathbf{H}$ with errors bounded by a factor ϵ. If u and v are any two feature vectors from $\mathbf{H}$, and u_p and v_p are the low-dimensional projected vectors, then;

$$(1-\epsilon)||u-v||^2 < ||u_p - v_p||^2 < (1+\epsilon)||u-v||^2 \tag{1}$$

1 holds provided that $p \geq \frac{4\ln(n)}{\epsilon^2/2 - \epsilon^3/3}$ (1). With $n = 1000$ for our training dataset, to preserve distances with a distortion factor of $\epsilon = .1$ requires ≥ 5920 dimensions. Thus we chose to project $\mathbf{H}$ to $\mathbf{P} \in \mathbb{R}^{n \times 5920}$ in instances where $m >> 5920$. Otherwise, $\mathbf{P} = \mathbf{H}$ and $p = m$. To find the mapping from $\mathbf{H}$ to $\mathbf{P}$ in the high dimensional case we used *sparse random projections* following Li et al. (27). The authors show a $\mathbf{P}$ satisfying 1 can be found by $\mathbf{P} = \mathbf{HR}$, where $\mathbf{R}$ is a sparse, $n \times p$ matrix, with i.i.d elements

$$r_{ji} = \begin{cases} \sqrt{\frac{\sqrt{m}}{p}} \text{ with prob. } \frac{1}{2\sqrt{m}} \\ 0 \text{ with prob. } 1 - \frac{1}{\sqrt{m}} \\ -\sqrt{\frac{\sqrt{m}}{p}} \text{ with prob. } \frac{1}{2\sqrt{m}} \end{cases} \tag{2}$$

LOOCV Ridge Regression Next, We used regularized (ridge) regression to predict voxel responses to images, $\mathbf{Y}$, from their associated (dimensionality-reduced) deep net features, $\mathbf{P}$. We first grouped our data into training and testing sets using a 50/50 split, such that $\mathbf{P}_{train} \in \mathbb{R}^{500 \times p}$ and $\mathbf{Y}_{train} \in \mathbb{R}^{500}$. In an abuse of notation, we'll refer to $\mathbf{P}_{train}$ and $\mathbf{Y}_{train}$ simply as $\mathbf{P}$ and $\mathbf{Y}$ when explaining the regression procedure. We aim to find a vector of coefficients, $\boldsymbol{\beta} \in \mathbb{R}^p$, that minimize $||\mathbf{P}\boldsymbol{\beta} - \mathbf{Y}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2$. λ is a hyper-parameter that constrains the norm of $\boldsymbol{\beta}$, which helps prevent overfitting in high-dimensional cases like ours, where $p > n$. To identify an optimal λ, we utilized a leave-one-out cross-validation procedure. We first standardized $\mathbf{Y}$ and the columns of $\mathbf{P}$ to have a mean of 0 and standard deviation of 1. Then for every image in our training set ($\forall i \in \{1 \dots 500\}$), Let $\mathbf{P}_{-i}$ and $\mathbf{Y}_{-i}$ denote $\mathbf{P}$ and $\mathbf{Y}$ with row i missing. One vector of coefficients per left-out image, $\boldsymbol{\beta}_i$, is calculated by;

$$\boldsymbol{\beta}_i = \left(\mathbf{P}'_{-i}\mathbf{P}_{-i} + \lambda I_p\right)^{-1} \mathbf{P}'_{-i}\mathbf{Y}_{-i} \tag{3}$$

Each $\boldsymbol{\beta}_i$ is then used to predict the voxel response of each left out image;

$$\hat{y}_i = \mathbf{P}_i\boldsymbol{\beta}_i, \quad \hat{\mathbf{Y}} = \{\hat{y}_i\}_{i=1}^{500} \tag{4}$$

$\hat{\mathbf{Y}}$ was computed over a logarithmic range of λ values, $\{10^j\}_{j=1}^{7}$, and the optimal λ was determined to be that with maximum pearson correlation between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$. We denote this maximum pearson correlation score for a given layer-to-voxel regression r_l^v, as it will be utilized latter in the analysis. The optimal λ was then used to calculate a general $\boldsymbol{\beta}$ following equation 3, using the full training set. We used the ***RidgeCV*** function from (34) to implement this cross-validated ridge regression, as its matrix algebraic implementation identifies each $\boldsymbol{\beta_i}$ in parallel, resulting in significant speedups (37). This procedure produced an individual vector of coefficients for each layer-to-voxel regression, $\boldsymbol{\beta}_l^v$.

Finally, we were interested in identifying model layers capable of encoding the voxel responses in entire ROI's in the brain. For each ROI (EVC, OTC, VWFA-1) we computed the average correlation score ($avg(r_l^v)$) for voxels in the ROI. We computed this average for each of a model's layers independently, and selected the layer, l, with the highest average score for each ROI. For each voxel in the ROI, we then used the previously computed $\boldsymbol{\beta}_l^v$ to predict its responses to the test set data, $\mathbf{Y}_{test} = \{y_i\}_{i=501}^{1000}$, as in equation 4. The correlation between the resultant $\hat{\mathbf{Y}}_{test}$ and $\mathbf{Y}_{test}$ corresponds to each models' ROI score as shown in Figure 1A.

A.4.2 REGRESSING IMAGE RATINGS (BEHAVIORAL DATA)

Our regression procedure for the behavioral data mostly follows that for the voxel data (see Appendix A.4.1 with some simplifications. Firstly rather than compute an optimal λ parameter for each model/layer/affect, we set $\lambda = 1e4$, a value that yielded the highest average LOOCV correlation score for layers in an AlexNet model that we subsequently exclude from the main analysis. Second, rather than split the data into train and test sets, we use all 900 images in the data set to define $\hat{\mathbf{Y}} \in \mathbb{R}^{900}$ with the LOOCV procedure described in equations 3 and 4. The model-wise scores reported in Figure 1B are each model's top $r(\mathbf{Y}, \hat{\mathbf{Y}})$ across all its layers.