Meta-Learning an In-Context Transformer Model of Human Higher Visual Cortex

Muquan Yu[1,2]   **Mu Nan**[1]   **Hossein Adeli**[3]   **Jacob S. Prince**[4]   **John A. Pyles**[5]

Leila Wehbe[6]   **Margaret M. Henderson**[6]   **Michael J. Tarr**[6]   **Andrew F. Luo**[†1]

[1] University of Hong Kong   [2] Chinese University of Hong Kong   [3] Columbia University
[4] Harvard University   [5] University of Washington   [6] Carnegie Mellon University

mqyu@link.cuhk.edu.hk   [†]Corresponding author: aluo@hku.hk

Abstract

Understanding functional representations within higher visual cortex is a fundamental question in computational neuroscience. While artificial neural networks pretrained on large-scale datasets exhibit striking representational alignment with human neural responses, learning image-computable models of visual cortex relies on individual-level, large-scale fMRI datasets. The necessity for expensive, time-intensive, and often impractical data acquisition limits the generalizability of encoders to new subjects and stimuli. **BraInCoRL** uses in-context learning to predict voxelwise neural responses from few-shot examples *without any additional finetuning* for novel subjects and stimuli. We leverage a transformer architecture that can flexibly condition on a variable number of in-context image stimuli, learning an inductive bias over multiple subjects. During training, we explicitly optimize the model for in-context learning. By jointly conditioning on image features and voxel activations, our model learns to directly generate better performing voxelwise models of higher visual cortex. We demonstrate that BraInCoRL consistently outperforms existing voxelwise encoder designs in a low-data regime when evaluated on entirely novel images, while also exhibiting strong test-time scaling behavior. The model also generalizes to an entirely new visual fMRI dataset, which uses different subjects and fMRI data acquisition parameters. Further, BraInCoRL facilitates better interpretability of neural signals in higher visual cortex by attending to semantically relevant stimuli. Finally, we show that our framework enables interpretable mappings from natural language queries to voxel selectivity. Our code and model weights are publicly available at https://github.com/leomqyu/BraInCoRL.

1   Introduction

Human visual cortex transforms raw sensory input into behaviorally-relevant representations of the world. While early visual areas are characterized by retinotopic organization and selective tuning to simple features such as edges and orientation gradients [1, 2, 3, 4, 5], higher-order visual areas demonstrate selectivity to more abstract semantics and categories. While this functional organization is largely consistent across individuals at a coarse scale, the spatial distribution and fine-grained semantic selectivity within visual cortex varies due to structural differences, developmental experience, and life-long learning [6, 7, 8, 9, 10, 11]. Such functional inter-subject differences pose a fundamental challenge in constructing generalizable models of higher visual cortex that can adapt to subject-specific neural organization without exhaustive data collection for every individual.

Recent advances in deep learning offer a promising avenue for addressing this challenge. Vision models pretrained on large-scale image datasets not only achieve strong object recognition performance, but also recapitulate hierarchical processing patterns observed in biological vision [12, 13, 14]. While these models may encapsulate some *universal principles of visual processing* [15], they do not inherently account for *individual* differences in cortical organization. To close the gap between artificial and biological systems, researchers have developed image-computable fMRI encoders – models that *predict brain activity from visual stimuli* [16]. These encoders typically regress image features onto voxelwise brain responses using subject-specific data, acting as computational probes of visual processing. Unfortunately, current approaches require many hours of costly fMRI scans per subject to fit these mappings – a prohibitive bottleneck for scalability to new populations, stimuli, and tasks, especially in clinical settings where collecting large amounts of data is difficult.

We bridge this gap with **BraInCoRL** (**Bra**in **In**-**Co**ntext **R**epresentation **L**earning), a transformer-based framework that meta-learns to predict subject-specific neural responses from provided examples. Inspired by language models that adapt to new tasks in-context, our approach treats voxel encoding as a function inference problem: given a handful of stimulus-response pairs from a new individual and novel stimuli, BraInCoRL constructs a voxelwise encoding model without any further training. By jointly optimizing for in-context learning across diverse subjects and stimuli, our model discovers shared functional principles of higher visual cortex that generalize to new subjects and stimuli represented by only a small amount of data. We illustrate the problem we are tackling in Figure 1.

We demonstrate that BraInCoRL: **(1)** Outperforms existing voxelwise encoder models in the low-data regime on novel visual stimuli while exhibiting strong generalization with increasing context. **(2)** Can generalize to new experiments with different scanning parameters. **(3)** Through analysis of attention values, learns to rely on images that are reflective of the category selected for in each region. **(4)** When paired with features from contrastive image-language models, facilitates zero-shot natural language–based characterization of cortical selectivity, enabling interpretable, finer-grained query-driven functional mapping.

2 Related work

Computational Encoding and Decoding Models for Visual Cortex. Computational modeling of neural data often involves two complementary approaches: encoding models that map from stimuli to neural activations, and decoding models that map from neural data to stimuli [16, 17, 18, 19, 20, 21, 22, 23, 24]. The development of both approaches has been facilitated by advances in machine learning models. For encoding models, the dominant approach is to combine pretrained deep feature extractors with linear voxelwise weights [25, 26, 27, 28, 29, 30]. More recent approaches have proposed to leverage transformers [31, 32] to learn the relationship between brain regions of a single subject. Most similar to our framework is the pioneering work by Adeli et al. [31] and Beliy & Wasserman et al. [33] which uses an auto-decoder based transformer network for multi-subject voxelwise encoding; However these approaches still require fine-tuning for novel subjects. More generally, encoders have been used to investigate the coding properties in higher-order visual areas [34, 35, 36, 37, 38, 39, 40, 41]. Encoders have been further combined with image-generation models [42, 43, 44, 45, 46, 47, 48, 49] or language-generation models [50, 51] to explore semantic selectivity. Recent progress on large generative models has enabled stimulus reconstruction from fMRI, EEG, and MEG signals for images [52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63], videos [64, 65, 66, 67, 68, 69, 70], and language/audio [71, 72, 73, 74, 75, 76, 77].

Representational Organization of Visual Cortex. Human visual cortex exhibits a hierarchical organization from primary visual to higher-order visual areas. The higher visual cortex is characterized by a tiling of semantically specialization. Approaches using functional localizers have identified category selective regions in higher visual that are responsive to faces [78, 79, 80, 81], places [82, 83], bodies [84], objects [85, 86], food [87, 88, 89], and words [90, 91]. While the spatial location of these broad category-selective regions are generally consistent across subjects [92], significant inter-individual variability exists in their anatomical location, spatial extent, and response profiles [6, 7, 8, 9, 93, 94, 95, 96, 97]. Accurately characterizing visual processing in higher-order visual areas necessitates subject-specific encoding models that capture individual diversity [98].

Meta-Learning and In-Context Learning. Our work builds upon meta-learning and in-context learning (ICL). Meta-learning trains models to "learn to learn" from a distribution of tasks, enabling
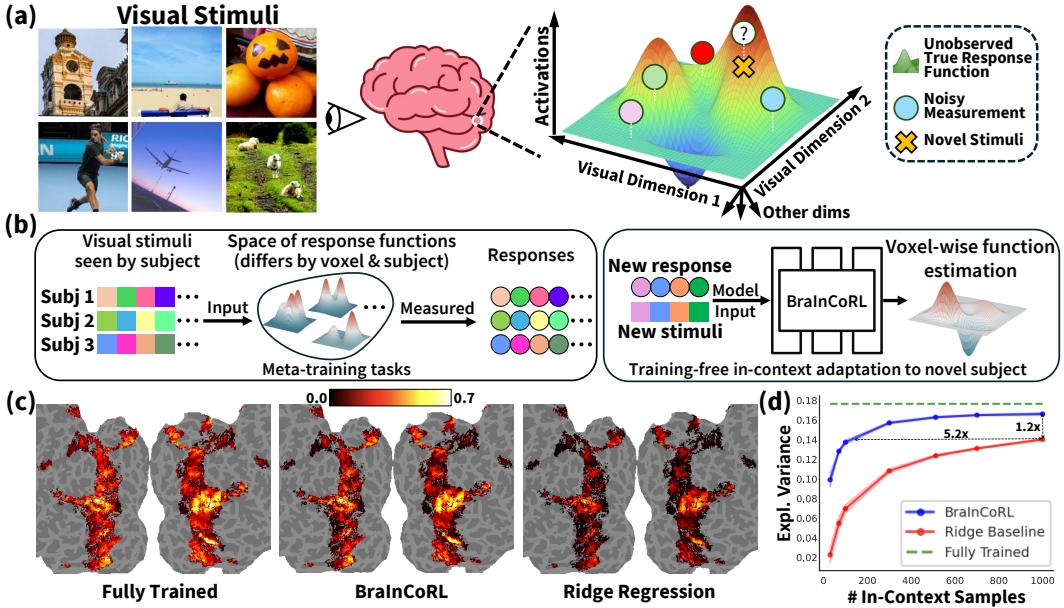
Figure 1: **BraInCoRL: Meta-Learning an In-Context Visual Cortex Encoder. (a)** The voxelwise brain encoding problem. For each voxel, there is a response function that maps from visual stimuli to voxel activation. In practice, we can only observe the noisy measurements from fMRI. The goal is to infer an image-computable function for each voxel to predict its activation. **(b)** BraInCoRL treats each voxel as a meta-learning task, and samples (image, response) pairs from multiple subjects. During testing, the model is conditioned on a small number of novel images and measurements from a new subject and directly outputs the function parameters. **(c)** From left to right, the explained variance from the full model trained on 9,000 images from one subject, BraInCoRL with only 100 in-context images from the new subject, and a baseline ridge regression also with 100 images (for this baseline, voxelwise regularization is determined using 5-way cross-validation). Our method achieves much higher data efficiency than baseline. **(d)** Explained variance as a function of in-context support set size. As the in-context support set size increases from 0 to 1,000, BraInCoRL steadily improves and approaches the fully trained reference model fit to converge on each subject's full 9,000-image training set, demonstrating high prediction accuracy and data efficiency.

quick adaptation to new tasks with few examples, via methods like meta-optimization [99, 100, 101] or metric-based approaches [102]. More recently, ICL has emerged as a powerful capability in Large Language Models [103, 104], where models adapt to new tasks at inference time solely based on examples provided in their prompt without any parameter updates [105, 106]. This has led to hypotheses that ICL is an emergent form of implicit meta-learning, where transformers effectively learn underlying learning algorithms during pre-training [107, 108]. Our goal is to learn the structure of functions that map between visual stimuli and voxelwise brain response. Our framework combines meta-training (across voxels and subjects) and in-context learning (across stimuli) to enable training free adaptation to novel subjects.

3 Methods

Our proposed framework leverages meta-learning and uses few-shot, in-context examples for voxelwise prediction of unseen stimuli (Figure 2). Critically, for unseen subjects, this approach does not require *any* additional finetuning. We achieve this by treating the mapping function from visual stimuli and the response of individual voxels as a set of meta-training tasks. This voxelwise approach is in line with higher-order visual areas being described by a multitude of functionally diverse voxels, which we randomly sample during training.

3.1 Motivation and Problem Definition

There is substantial inter-subject anatomical and functional variability in the higher visual cortex among humans. Consequently, while one can learn per-subject image-computable encoders that map image features to brain responses with high predictive accuracy, these models require large amounts
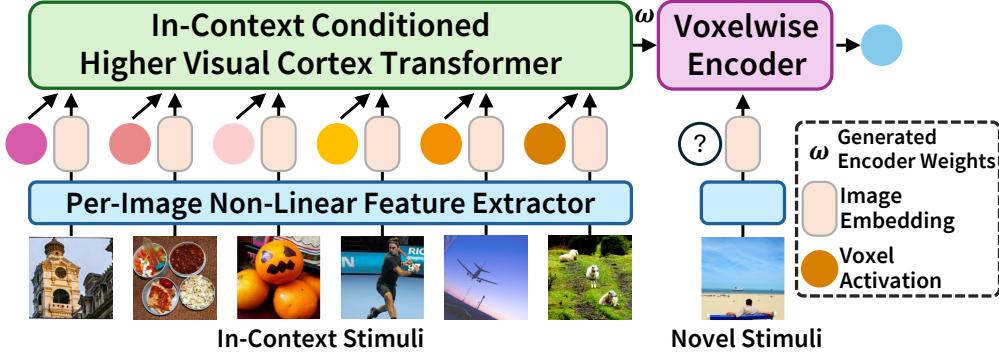
Figure 2: **Architecture of the In-Context Voxelwise Encoder (BraInCoRL).** **(1)** A pretrained feature extractor converts visual stimuli into vector embeddings. **(2)** A higher visual cortex transformer integrates these embeddings with voxel activations to learn context-specific features and generates hyperweights for a subsequent voxelwise encoder backbone. **(3)** The voxelwise encoder, conditioned on the hyperweights, predicts voxel responses for novel stimuli.

of within-subject data and do not exploit information across subjects. To account for this variability across individuals, we design our framework to treat individual voxels as the fundamental unit of modeling. Importantly, our method does not assume any overlap in stimuli across subjects, yet still enables us to take advantage of multi-subject training data.

We formalize this problem by assuming an image j is represented as RGB tensor $\mathcal{I}_j \in \mathbb{R}^{H \times W \times 3}$. Given an image j and a human subject k, there is a 1D array of voxel activations (beta values) from higher visual cortex: $(\beta_1, \beta_2, ..., \beta_{N_k})_j = B_{j,k} \in \mathbb{R}^{1 \times N_k}$, where the number of voxels N will differ between subjects.

Given a new subject not seen during training, we have a small set of n seen images $(\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_n)$ and measured brain responses $(B_1, B_2, ..., B_n)$ for this new subject. *Our goal is to estimate the brain response to an arbitrary new image $\mathcal{I}_{\text{novel}}$.*

3.2 Meta-Learning an In-Context Transformer

Image-computable encoders that map from images to brain responses for a single subject k are typically considered as a function $f_k(\mathcal{I}) \Rightarrow B$, and jointly model the entire visual cortex. While powerful, this approach cannot be easily extended to the multi-subject scenario, where test-time individuals may have functional and anatomical differences that are not known during training. In contrast, BraInCoRL considers each voxel v to have a unique and unobserved visual response function $f_{k,v}(\mathcal{I}) \Rightarrow \beta_v$. Voxels can be from many different subjects. During training, we consider each voxel's response function to be a meta-training task, where each task is effectively specified by input images and voxel response pairs. In order to facilitate training-free adaptation on new subjects, we utilize in-context learning across stimuli enabled by a transformer backbone.

For a single voxel we define a support set of p images and neural responses $\{(x_1, \beta_1), (x_2, \beta_2), ..., (x_p, \beta_p)\}$, where $x_i \in \mathbb{R}^m$ is the image embedding vector extracted by a frozen image feature extractor $\phi(\cdot)$, i.e., $x_i = \phi(I_i)$, and $\beta_i \in \mathbb{R}$ is the voxel's response observed for image I_i. Each pair is concatenated to form context tokens $c_i = [x_i; \beta_i]$, and the full context is defined as $\{c_1, ..., c_p\}$. Unlike traditional in-context inference in large language models, where there is a query concatenated to the end of the context, we avoid expensive per-voxel inference by directly generating the parameters for the voxelwise visual response function. During training, we optimize the BraInCoRL transformer T with parameters θ such that it outputs voxel response function f with parameters ω:

$$\omega = T_\theta(c_1, c_2, ..., c_p) \tag{1}$$

$$\hat{\beta} = f_\omega(\mathcal{I}) \tag{2}$$

Since T and f are differentiable, we optimize θ to maximize the likelihood of observing β given $\mathcal{I}$:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(I_q, \beta_q)} \| f_\omega(\mathcal{I}) - \beta_{\text{True}} \|_2^2 \tag{3}$$

In practice, we use mean-squared-error and gradient based mini-batch optimization.

4

Table 1: **Voxelwise performance across five category-selective regions.** Explained variance is shown for our in-context model ("BraInCoRL") that uses just 100 in-context images, the fully trained reference model fit to converge on each subject's full 9,000-image training set ("Fully Trained"), and within-subject ridge regression baselines (100, 300 within-subject test images), plus the FsAverage map averages over other subjects. Our model outperforms both subject-wise and anatomical baselines, and demonstrates strong data-efficiency.

	Faces		Places		Bodies		Words		Food		Mean	
	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fully Trained	0.19	0.16	0.20	0.27	0.28	0.24	0.11	0.11	0.16	0.17	0.18	0.19
Ridge-100	0.10	0.07	0.08	0.14	0.16	0.12	0.02	0.03	0.05	0.07	0.07	0.08
Ridge-300	0.13	0.10	0.13	0.20	0.22	0.16	0.06	0.06	0.10	0.11	0.11	0.12
FsAverage map	0.13	0.06	0.11	0.19	0.09	0.08	0.06	0.03	0.14	0.18	0.08	0.06
BraInCoRL-100	**0.16**	**0.13**	**0.16**	**0.23**	**0.25**	**0.21**	**0.07**	**0.08**	**0.12**	**0.13**	**0.13**	**0.15**

3.3   Test-time Context Scaling

At test time, when we encounter a new subject, we assume we have access to a small set of novel images and the corresponding brain responses – we want to predict a voxelwise encoder. While our voxelwise parameterization successfully resolves the challenge of unknown test-time anatomy and geometry, the challenge of unknown test-time context size remains. Unlike transformers in language models, where the output is dependent on the order of the samples, we want our model to be explicitly invariant to input order. In order to facilitate variable length context, we utilize logit scaling [109, 110, 111]. Assuming a query/key (q, k) with d_k features and a length l context:

$$\alpha_{\text{orig}} = \frac{q \cdot k}{\sqrt{d_k}}; \quad \alpha_{\text{scaled}} = \frac{\log{(l)} \cdot q \cdot k}{\sqrt{d_k}} \qquad (4)$$

We find this method effectively enables context scaling when trained with variable length context. While the hypernetwork could, in principle, parameterize any class of neural encoders (e.g., MLPs, convolution, attention layers), prior studies utilizing brain data have largely used linear parameterizations that map from deep network features to voxel responses [13, 14], and find that such a choice offers high performance and interpretability. Given features $x \in \mathbb{R}^{1 \times q}$ from a pretrained neural network $x = \phi(\mathcal{I})$, we adopt the same setup and predict the final voxel response:

$$\hat{\beta} = f(\phi(\mathcal{I}); \omega) = \texttt{matmul}(x, \omega) \qquad (5)$$

A more detailed description on the test-time context scaling technique is provided in Appendix A.4.

4   Experiments

We utilize BraInCoRL to generate encoder weights in a low-data regime. We start by describing our experiment setup. First, we evaluate the effectiveness of BraInCoRL on **novel subjects** where there is *zero overlap between the training dataset and the evaluated subject's in-context stimulus*. We also evaluate our framework where data from novel subjects are collected from a **completely different scanner and protocol**. Second, we explore the attention pattern across stimuli for different ROIs, and perform ablations to explore the need for test-time in-context stimulus diversity. Third, we show that our method enables natural language characterizations of the visual cortex using very little data.

4.1   Setup

Dataset.   We primarily perform experiments with the Natural Scenes Dataset (NSD) [112], but then validate with the BOLD5000 dataset [113]. Both are large-scale neural datasets: NSD is the largest 7T fMRI image viewing dataset available, where eight subjects each viewed $\sim 10,000$ images; BOLD5000 is a 3T dataset, where four subjects each viewed ~ 5000 images. In NSD each image was viewed up to three times, while in BOLD5000 only a subset of images were viewed up to four times. For NSD, of the eight subjects, four subjects (S1, S2, S5, S7) completed scanning and are the focus of our analysis in the main paper. The results of other subjects are presented in the supplemental
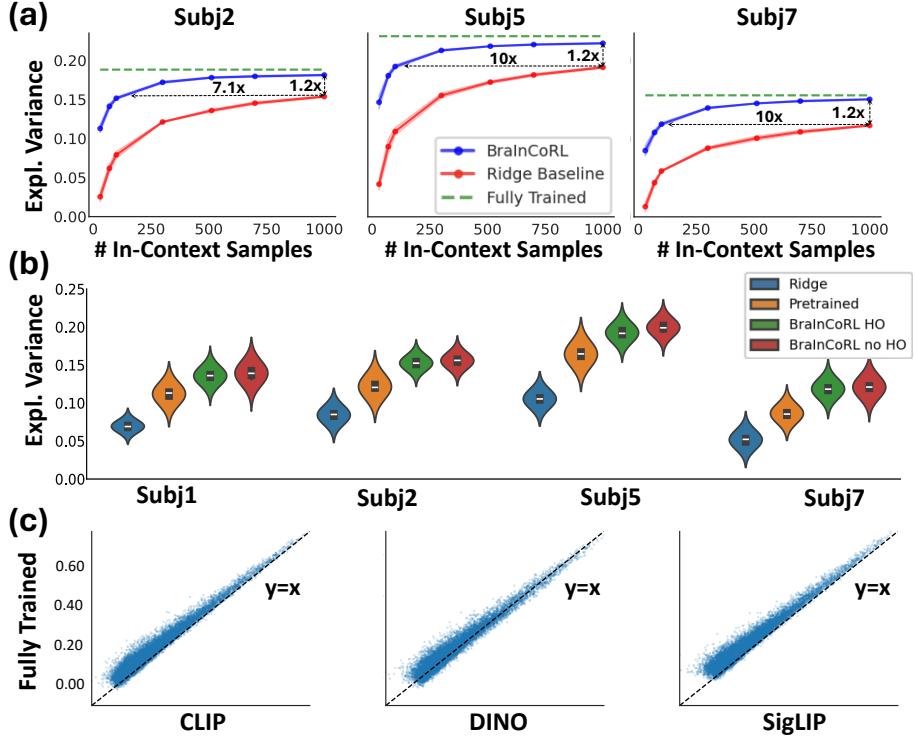
Figure 3: **Evaluation on NSD. (a)** Prediction explained variance of BraInCoRL improves on novel subjects with larger in-context support set size, outperforming within-subject ridge regression and approaching the fully trained reference model fit on each subject's full 9,000-image training set, using far less data. **(b)** Ablation (100 support images) comparing BraInCoRL variants: the original model trained while holding out the novel subject's 9,000 test-time support images ("HO"), a BraInCoRL model trained without this holdout ("no HO"), and a pretraining-only BraInCoRL model, alongside the within-subject ridge baseline. Results show that finetuning with real fMRI data improves performance, and holding out the test subject's image data does not hinder generalization. **(c)** Voxelwise explained variance from BraInCoRL (100 images) is strongly correlated with fully trained reference models across different visual encoder backbones. Note that the y-axis represents explained variance of the fully trained model (9,000 images), while x-axis represents explained variance of BraInCoRL.

materials. For each subject, $\sim 9,000$ images are unique to each other, while $\sim 1,000$ are viewed by all eight subjects. The NSD stimuli were sourced from the MS COCO dataset (as were a subset of the BOLD5000 stimuli). Unless otherwise noted, we perform our analysis in subject native volume space (`func1pt8mm`) for NSD, where the voxelwise betas are z-scored within each session then averaged across repeats of the same stimulus. In order to rigorously evaluate BraInCoRL for a given subject, we use the $3 \times 9,000$ unique images viewed by the other three subjects as the meta-training data. During the ROI-wise evaluation for NSD, we follow prior work [48] and apply a t-statistic cutoff of $t > 2$ by leveraging independent functional localizer data provided with the dataset to threshold the originally broad definitions. During quantitative evaluation, we follow prior work [114] and apply a voxel quality cutoff of `ncsnr` > 0.2. For BOLD5000, we use a model trained on four NSD subjects (S1, S2, S5, S7). Following the suggestion of BOLD5000 authors, we only model stimuli with 4 repeats and apply a cutoff of `ncsnr` > 0.3. Voxel responses are averaged across repeats.

Training and evaluation. Our training process takes inspiration from LLM based training setups, and we adopt a three stage process – pretraining, context extension, and supervised fine tuning. In the pretraining stage, we use an analysis-by-synthesis approach without relying on any (real) subject data. We artificially construct a large number of voxels with synthesized weights. We derive synthetic voxel responses with normally distributed noise using these synthesized weights and train our model using a fixed context size of 500. In the second stage, we randomly sample the context size from `Uniform`$(30, 500)$ which allows the model to acquire length robustness. Finally, in the finetuning stage, the model is trained on real fMRI data using the subject-specific beta values,
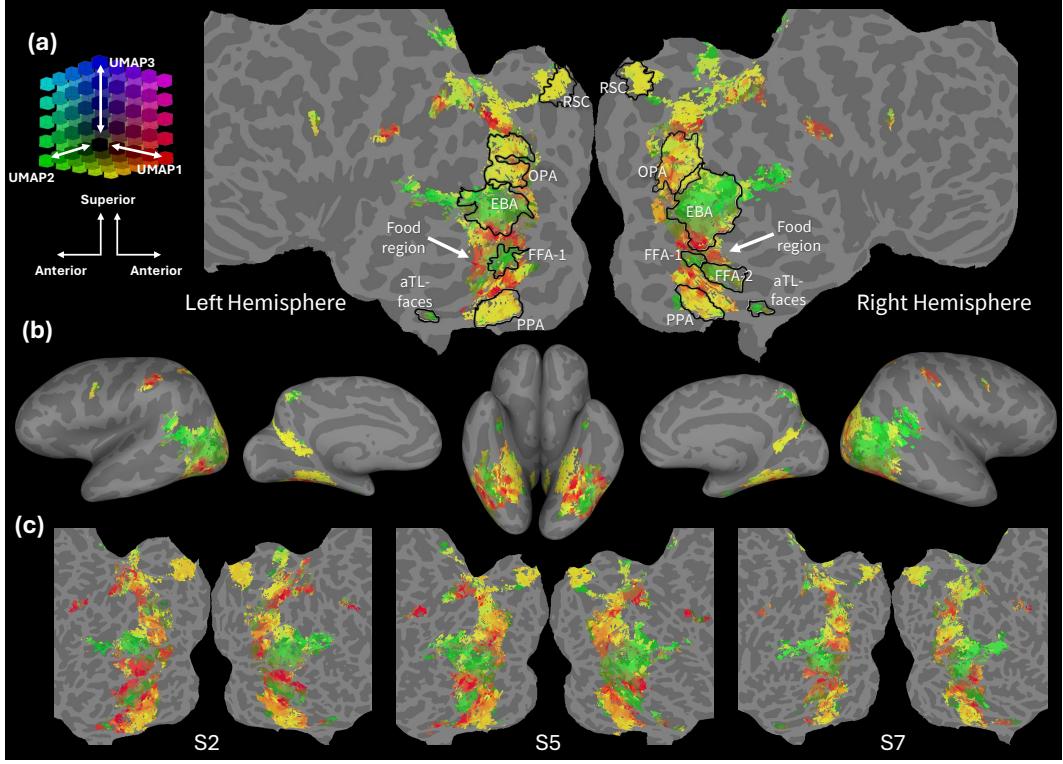
Figure 4: **UMAP visualization of predicted response weights.** We apply UMAP to BraInCoRL
-predicted voxelwise weights (100 support images) and show: (a) a flatmap for S1 with ROI outlines,
(b) the same projection on an inflated surface, and (c) flatmaps for S2, S5, and S7. Color-coded
clusters align with body/face regions (EBA, FFA/aTL-faces), place regions (RSC, OPA, PPA), and
food regions (in red).

enabling adaptation to biologically grounded neural responses. All our evaluation experiments are
performed on **novel subjects** that are unseen by the model during training, with exception of (Fig-
ure 3b) no heldout ("no HO") ablation study.

4.2 Effectiveness of In-Context Higher Visual Cortex Encoders

On NSD. In this experiment, we evaluated BraInCoRL using each subject's 9,000 unique images
as the in-context support set and the shared 1,000 images as the test set. For each evaluation, we
randomly sampled training images from the subject-specific in-context support set and test images
from the shared test set. Explained variance statistics averaged over category-selective ROIs are
reported in Table 1. We compare BraInCoRL using just 100 test images against within-subject
ridge regression baseline trained on 100 and 300 support-set images of the test-subject, with the
regularization strength selected via 5-fold cross-validation over $[10^{-3}, 10^{-2}, \ldots, 10^{8}]$. Remarkably,
BraInCoRL with only 100 images nearly matches the performance of the fully supervised reference
model that is trained by gradient descent on each subject's entire within-subject support-set of 9,000
images until convergence. We also evaluate an anatomical FsAverage baseline which aligns each
training subject's anatomy to a common template and projects the average response onto novel
subjects for prediction. While this baseline benefits from a strong anatomical prior, it is outperformed
by BraInCoRL, which directly adapts to each subject's unique neural responses with higher efficiency.

To evaluate test-time behavior, we assess how performance scales with increasing in-context support
size. BraInCoRL consistently outperforms within-subject ridge regression and more efficiently
approaches the fully trained reference model (Figure 1c for subject 1 and Figure 3a for subject 2, 5,
7). Moreover, we conduct ablations by evaluating a BraInCoRL model trained without holding out
the test subject's support images and a BraInCoRL model with only pretraining. Results confirm
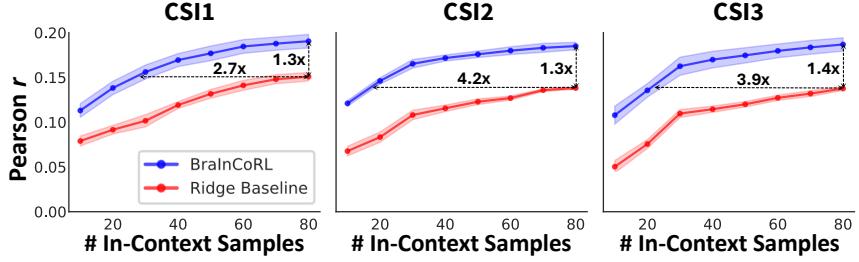
Figure 5: **Evaluation on BOLD5000.** We evaluate BraInCoRL on the BOLD5000 dataset, which was collected using a different scanner than NSD. For varying in-context support set sizes, we report voxelwise Pearson correlation between predicted and true responses for both BraInCoRL and wi...

Figure 6: **Top contributing support images for each category-selective region in S1.** For each of the five category-selective regions, we select the in-context support images with the highest attention weights in BraInCoRL's final attention layer for voxels in that region. We visualize the top 5 contributing images for the place, word, face and body regions, and the top 10 for the food region.

that finetuning with real neural data boosts performance and that BraInCoRL can generalize well to previously unseen images without overfitting (Figure 3b). Additionally, we observe high voxelwise explained variance correlation between BraInCoRL and the fully trained reference model across multiple backbones (Figure 3c). Finally, we apply UMAP to the BraInCoRL predicted response-function weights, revealing clear semantic clustering across higher visual areas (Figure 4) that correspond with known visual regions.

On BOLD5000. We validate generalization on the BOLD5000 dataset in Figure 5. BOLD5000 has many differences with NSD and represents the challenge of cross-site generalization that is the main objective of our method. BOLD5000 was collected on a 3T scanner with a different stimulus presentation time, a slow-event related trial structure (10s inter-trial interval), different images and image datasets, a different voxel size (2mm isotropic), and different subjects. BraInCoRL achieves higher voxelwise Pearson correlations than within-subject ridge regression. Moreover, results remain consistent across different subjects, demonstrating the robustness and reliability of our method.

4.3 Semantic Discovery through Text–Image Embedding Alignment

To better understand how BraInCoRL leverages in-context examples, we analyze its internal attention mechanisms to identify images that strongly influence voxel predictions in category-selective regions. In Figure 6, we examine attention weights from BraInCoRL's final attention layer to determine the top-contributing images for each cortical region. The visualized images with the highest attention scores closely align with known semantic preferences of the respective cortical regions.

However, Figure 7 reveals a counterintuitive finding regarding the semantic specificity of in-context support sets. We systematically vary the specificity of the 100-image sets provided to the model, ranging from highly relevant to random selections. Selections are determined via the first text-prompt in each category (see Appendix). We observe that randomly selected images lead to better predictive performance compared to sets composed solely of highly relevant images. This suggests that overly specific context sets may limit the generalization capabilities of the encoder system, and diverse, less semantically constrained images provide richer context for learning robust voxel representations.

8
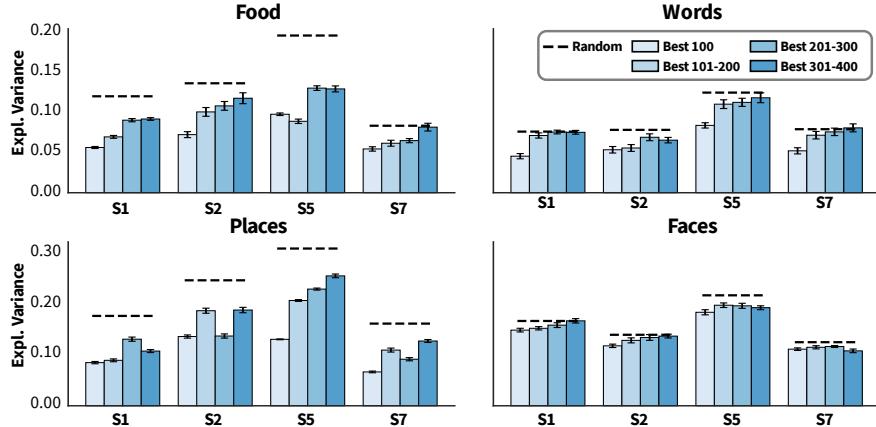
Figure 7: **Impact of support-set specificity on category-selective ROI encoding performance on NSD.** We construct in-context support sets of 100 images based on descending semantic relevance for each ROI (tiers: 1–100, 101–200, 201–300, 301–400) and compare them with randomly sampled sets of equal size. Mean explained variance in the target category-selective ROIs increases as semantic specificity decreases, with all curated sets performing worse than random sampling. This suggests that overly specific support sets hinder generalization in voxelwise encoding. This pattern echoes prior findings on diverse stimuli contributing to better encoders [13].
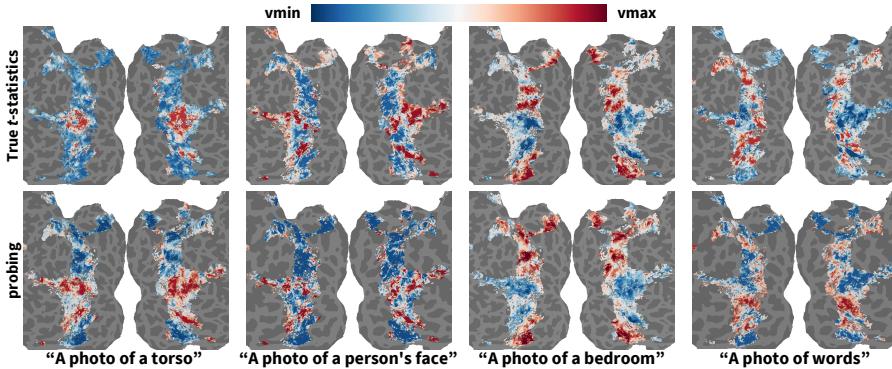

Figure 8: **Predicting cortical responses from natural language prompts.** For each semantic category, we convert a natural language prompt into a CLIP text embedding, project it into the image feature space, and use BraInCoRL to predict the corresponding voxel activation map. The predicted activations align closely with true t-statistic of category-selective regions (derived from fMRI functional localizer experiments), illustrating the potential for efficient, language-driven functional mapping of visual cortex.

Table 2: **Voxelwise prompt classification accuracy.** Each cell shows the percentage of voxels in a given category selective region (columns) whose peak predicted activation was elicited by a specific semantic prompt (rows, see Appendix). Using only 100 support images, BraInCoRL effectively localizes category-selective regions with high data efficiency.

	Bodies		Faces		Places		Food		Words	
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**
Bodies	**0.63**	**0.54**	0.30	0.16	0.05	0.03	0.15	0.19	**0.43**	0.17
Faces	0.30	0.25	**0.60**	**0.56**	0.05	0.01	0.07	0.04	0.15	0.16
Places	0.02	0.09	0.02	0.05	**0.81**	**0.88**	0.10	0.07	0.05	0.10
Food	0.04	0.10	0.08	0.18	0.08	0.06	**0.66**	**0.64**	0.31	**0.45**
Words	0.01	0.03	0.00	0.04	0.01	0.02	0.02	0.05	0.05	0.12

4.4 Characterizing higher visual cortex with text embeddings to images

In this experiment, we investigate the capability of BraInCoRL to enable interpretable, query-driven functional mapping using natural language prompts.

In Figure 8, we demonstrate that natural language prompts can be effectively mapped to predicted voxel activations. For each category selective region, we convert the corresponding natural language prompt into a CLIP embedding and project it into the image feature space to directly predict voxel activations. The resulting activation maps closely match expected t-statistics, reflecting BraInCoRL's ability to support intuitive, language-driven cortical queries.

In the second analysis (Table 2), we quantitatively assess the accuracy of prompt-driven activation predictions. We measure the fraction of voxels within each category-selective region whose peak predicted activation aligns with the category indicated by the natural language query. Results confirm that BraInCoRL paired with language embeddings achieves a high level of alignment between predicted voxel selectivity and query semantics across multiple subjects. The predictions for word-selective voxels were notably less accurate. We hypothesize this discrepancy arises from the developmental and experiential variability inherent to the formation of word-selective regions, as these areas form predominantly through individualized learning experiences, such as reading and linguistic exposure, leading to greater inter-subject variability.

5 Discussion

Limitations and Future Work.   Here we have shown that meta-learning an in-context model of higher visual cortex can yield high performance and strong data efficiency gains, outperforming anatomical (FsAverage) and subject-wise baselines on novel subjects. Our work currently focuses on static natural images, and extensions to dynamic stimuli would likely require a rethinking of the encoder backbone and network structure. Further, while we show strong generalization across scanners and utilize the largest fMRI dataset that is NSD, there may still be limitations in dataset diversity [115]. Collection of larger and more diverse fMRI datasets will help mitigate this issue.

Conclusion.   We introduce a foundation model that serves as an fMRI encoder, mapping from natural images to voxelwise activations. We demonstrate that our method can adapt without any finetuning to new stimuli, subjects, scanners, and scanning protocols. Our model achieves this by meta-learning across voxels from different subjects, and performing in-context learning across stimuli. Our approach has significant data-efficiency and performance gains over baseline methods, and has the potential to help understand cortical structure in data-constrained environments.

References

[1] Stephen W Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1):37–68, 1953.

[2] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

[3] Joseph J Atick and A Norman Redlich. What does the retina know about natural scenes? *Neural computation*, 4(2):196–210, 1992.

[4] Jan Verweij, Eric P Hornstein, and Julie L Schnapf. Surround antagonism in macaque cone photoreceptors. *Journal of Neuroscience*, 23(32):10249–10257, 2003.

[5] Kalanit Grill-Spector and Rafael Malach. The human visual cortex. *Annu. Rev. Neurosci.*, 27(1): 649–677, 2004.

[6] Michael J Tarr and Isabel Gauthier. FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature neuroscience*, 3(8):764–769, 2000.

[7] Isabel Gauthier, Pawel Skudlarski, John C Gore, and Adam W Anderson. Expertise for cars and birds recruits brain areas involved in face recognition. *Nature neuroscience*, 3(2):191–197, 2000.

[8] Roel M Willems, Marius V Peelen, and Peter Hagoort. Cerebral lateralization of face-selective and body-selective visual areas depends on handedness. *Cerebral cortex*, 20(7):1719–1725, 2010.

[9] Qing Cai, Lise Van der Haegen, and Marc Brysbaert. Complementary hemispheric specialization for language production and visuospatial attention. *Proceedings of the National Academy of Sciences*, 110(4):E322–E330, 2013.

[10] Philippe Pinel, Christophe Lalanne, Thomas Bourgeron, Fabien Fauchereau, Cyril Poupon, Eric Artiges, Denis Le Bihan, Ghislaine Dehaene-Lambertz, and Stanislas Dehaene. Genetic and environmental influences on the visual word form and fusiform face areas. *Cerebral Cortex*, 25 (9):2478–2493, 2015.

[11] Zeynep M Saygin, David E Osher, Elizabeth S Norton, Deanna A Youssoufian, Sara D Beach, Jenelle Feather, Nadine Gaab, John DE Gabrieli, and Nancy Kanwisher. Connectivity precedes function in the development of the visual word form area. *Nature neuroscience*, 19(9):1250–1255, 2016.

[12] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[13] Aria Yuan Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nat Mach Intell*, 5:1415–1426, 2023.

[14] Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pages 2022–03, 2022.

[15] Zirui Chen and Michael F. Bonner. Universal dimensions of visual representation, 2024.

[16] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[17] Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005.

[18] Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430, 2006.

[19] Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, 2019.

[20] Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018.

[21] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019.

[22] Ziqi Ren, Jie Li, Xuetong Xue, Xin Li, Fan Yang, Zhicheng Jiao, and Xinbo Gao. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228:117602, 2021.

[23] Yuqin Dai, Zhouheng Yao, Chunfeng Song, Qihao Zheng, Weijian Mai, Kunyu Peng, Shuai Lu, Wanli Ouyang, Jian Yang, and Jiamin Wu. Mindaligner: Explicit brain functional alignment for cross-subject visual decoding from limited fmri data, 2025. URL https://arxiv.org/abs/2502.05034.

[24] Alessandro T Gifford, Benjamin Lahner, Pablo Oyarzo, Aude Oliva, Gemma Roig, and Radoslaw M Cichy. What opportunities do large-scale visual neural datasets offer to the vision sciences community? *Journal of Vision*, 24(10):152–152, 2024.

[25] Serge O Dumoulin and Brian A Wandell. Population receptive field estimates in human visual cortex. *Neuroimage*, 39(2):647–660, 2008.

[26] Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014, 2015.

[27] David Klindt, Alexander S Ecker, Thomas Euler, and Matthias Bethge. Neural system identification for large populations separating "what" and "where". *Advances in neural information processing systems*, 30, 2017.

[28] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152: 184–194, 2017.

[29] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12):4136–4160, 2018.

[30] Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254:119121, 2022.

[31] Hossein Adeli, Sun Minni, and Nikolaus Kriegeskorte. Predicting brain activity using transformers. *bioRxiv*, pages 2023–08, 2023.

[32] Guangyin Bao, Qi Zhang, Zixuan Gong, Zhuojia Wu, and Duoqian Miao. Mindsimulator: Exploring brain concept localization via synthetic fmri. *arXiv preprint arXiv:2503.02351*, 2025.

[33] Roman Beliy, Navve Wasserman, Amit Zalcher, and Michal Irani. The wisdom of a crowd of brains: A universal brain encoder. *arXiv preprint arXiv:2406.12179*, 2024.

[34] Meenakshi Khosla and Leila Wehbe. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, pages 2022–03, 2022.

[35] Meenakshi Khosla, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. Characterizing the ventral visual stream with response-optimized neural encoding models. *Advances in Neural Information Processing Systems*, 35:9389–9402, 2022.

[36] Cory Efird, Alex Murphy, Joel Zylberberg, and Alona Fyshe. What's the opposite of a face? finding shared decodable concepts and their negations in the brain. *arXiv e-prints*, pages arXiv–2405, 2024.

[37] Huzheng Yang, James Gee, and Jianbo Shi. Brain decodes deep nets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23030–23040, 2024.

[38] Huzheng Yang, James Gee, and Jianbo Shi. Alignedcut: Visual concepts discovery on brain-guided universal feature space. *arXiv preprint arXiv:2406.18344*, 2024.

[39] Andrew F Luo, Jacob Yeung, Rushikesh Zawar, Shaurya Dewan, Margaret M Henderson, Leila Wehbe, and Michael J Tarr. Brain mapping with dense features: Grounding cortical semantic selectivity in natural images with vision transformers. *arXiv preprint arXiv:2410.05266*, 2024.

[40] Gabriel H. Sarch, Michael J. Tarr, Katerina Fragkiadaki, and Leila Wehbe. Brain dissection: fmri-trained networks reveal spatial selectivity in the processing of natural images. *bioRxiv*, 2023. doi: 10.1101/2023.05.29.542635. URL https://www.biorxiv.org/content/early/2023/11/20/2023.05.29.542635.

[41] Alexander Lappe, Anna Bognár, Ghazaleh Ghamkahri Nejad, Albert Mukovskiy, Lucas Martini, Martin Giese, and Rufin Vogels. Parallel backpropagation for shared-feature visualization. *Advances in Neural Information Processing Systems*, 37:22993–23012, 2024.

[42] Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22 (12):2060–2065, 2019.

[43] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

[44] Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009, 2019.

[45] N Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1):5540, 2021.

[46] Zijin Gu, Keith Wakefield Jamison, Meenakshi Khosla, Emily J Allen, Yihan Wu, Ghislain St-Yves, Thomas Naselaris, Kendrick Kay, Mert R Sabuncu, and Amy Kuceyeski. NeuroGen: activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247:118812, 2022.

[47] Paweł A Pierzchlewicz, Konstantin F Willeke, Arne F Nix, Pavithra Elumalai, Kelli Restivo, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Katrin Franke, et al. Energy guided diffusion for generating neurally exciting images, 2023.

[48] Andrew F Luo, Margaret M Henderson, Leila Wehbe, and Michael J Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *arXiv preprint arXiv:2306.03089*, 2023.

[49] Diego García Cerdas, Christina Sartzetaki, Magnus Petersen, Gemma Roig, Pascal Mettes, and Iris Groen. Brainactiv: Identifying visuo-semantic properties driving cortical selectivity using diffusion-based image manipulation. *bioRxiv*, pages 2024–10, 2024.

[50] Andrew Luo, Margaret Marie Henderson, Michael J. Tarr, and Leila Wehbe. Brainscuba: Fine-grained natural language captions of visual cortex selectivity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mQYHXUUTkU.

[51] Takuya Matsuyama, Shinji Nishimoto, and Yu Takagi. Lavca: Llm-assisted visual cortex captioning. *arXiv preprint arXiv:2502.13606*, 2025.

[52] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*, pages 2022–11, 2022.

[53] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023.

[54] Yizhuo Lu, Changde Du, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. *arXiv preprint arXiv:2303.14139*, 2023.

[55] Furkan Ozcelik and Rufin VanRullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. *arXiv preprint arXiv:2303.05334*, 2023.

[56] Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*, 2022.

[57] Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*, 2023.

[58] Yulong Liu, Yongqiang Ma, Wei Zhou, Guibo Zhu, and Nanning Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding from fmri. *arXiv preprint arXiv:2302.12971*, 2023.

[59] Weijian Mai and Zhijun Zhang. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv preprint arXiv:2308.07428*, 2023.

[60] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

[61] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*, 2023.

[62] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024.

[63] Zhanqiang Guo, Jiamin Wu, Yonghao Song, Jiahui Bu, Weijian Mai, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. Neuro-3d: Towards 3d visual decoding from eeg signals, 2024. URL https://arxiv.org/abs/2411.12248.

[64] Yu Zhu, Bo Lei, Chunfeng Song, Wanli Ouyang, Shan Yu, and Tiejun Huang. Multi-modal latent variables for cross-individual primary visual cortex modeling and analysis, 2024. URL https://arxiv.org/abs/2412.14536.

[65] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023.

[66] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36: 24841–24858, 2023.

[67] Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, Ke Liu, and Yu Zhang. Neuroclips: Towards high-fidelity and smooth fMRI-to-video reconstruction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=8qu52Fl1Dt.

[68] Jacob Yeung, Andrew F Luo, Gabriel Sarch, Margaret M Henderson, Deva Ramanan, and Michael J Tarr. Neural representations of dynamic visual stimuli. *arXiv preprint arXiv:2406.02659*, 2024.

[69] Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from eeg signals. *Advances in Neural Information Processing Systems*, 37:72245–72273, 2024.

[70] Camilo Fosco, Benjamin Lahner, Bowen Pan, Alex Andonian, Emilie Josephs, Alex Lascelles, and Aude Oliva. Brain netflix: Scaling data to reconstruct videos from brain signals. In *European Conference on Computer Vision*, pages 457–474. Springer, 2024.

[71] Brian N Pasley, Stephen V David, Nima Mesgarani, Adeen Flinker, Shihab A Shamma, Nathan E Crone, Robert T Knight, and Edward F Chang. Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1):e1001251, 2012.

[72] Gaël Varoquaux, Pradeep Reddy Raamana, Denis A Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017.

[73] Ludovic Bellier, Anaïs Llorens, Déborah Marciano, Aysegul Gunduz, Gerwin Schalk, Peter Brunner, and Robert T Knight. Music can be reconstructed from human auditory cortex activity using nonlinear decoding models. *PLoS biology*, 21(8):e3002176, 2023.

[74] Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. *arXiv preprint arXiv:2311.04664*, 2023.

[75] Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are eeg-to-text models working? *arXiv preprint arXiv:2405.06459*, 2024.

[76] Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

[77] Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976): 1037–1046, 2023.

[78] Justine Sergent, Shinsuke Ohta, and Brennan Macdonald. Functional neuroanatomy of face and object processing: a positron emission tomography study. *Brain*, 115(1):15–36, 1992.

[79] Truett Allison, Gregory McCarthy, Anna Nobre, Aina Puce, and Aysenil Belger. Human extrastriate visual cortex and the perception of faces, words, numbers, and colors. *Cerebral cortex*, 4(5):544–554, 1994.

[80] Gregory McCarthy, Aina Puce, John C Gore, and Truett Allison. Face-specific processing in the human fusiform gyrus. *Journal of cognitive neuroscience*, 9(5):605–610, 1997.

[81] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11): 4302–4311, 1997.

[82] Geoffrey K Aguirre, John A Detre, David C Alsop, and Mark D'Esposito. The parahippocampus subserves topographical learning in man. *Cerebral cortex*, 6(6):823–829, 1996.

[83] Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

[84] Paul E Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.

[85] Kalanit Grill-Spector. The neural basis of object perception. *Current opinion in neurobiology*, 13(2):159–166, 2003.

[86] Rafael Malach, JB Reppas, RR Benson, KK Kwong, H Jiang, WA Kennedy, PJ Ledden, TJ Brady, BR Rosen, and RB Tootell. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, 92(18):8135–8139, 1995.

[87] Meenakshi Khosla, N. Apurva Ratan Murty, and Nancy Kanwisher. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32:1–13, 2022.

[88] Ian ML Pennock, Chris Racey, Emily J Allen, Yihan Wu, Thomas Naselaris, Kendrick N Kay, Anna Franklin, and Jenny M Bosten. Color-biased regions in the ventral visual pathway are food selective. *Current Biology*, 33(1):134–146, 2023.

[89] Nidhi Jain, Aria Wang, Margaret M. Henderson, Ruogu Lin, Jacob S. Prince, Michael J. Tarr, and Leila Wehbe. Selectivity for food in human ventral visual cortex. *Communications Biology 2023 6:1*, 6:1–14, 2 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04546-2.

[90] Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff, and François Michel. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2):291–307, 2000.

[91] Stanislas Dehaene, Lionel Naccache, Laurent Cohen, Denis Le Bihan, Jean-François Mangin, Jean-Baptiste Poline, and Denis Rivière. Cerebral mechanisms of word masking and unconscious repetition priming. *Nature neuroscience*, 4(7):752–758, 2001.

[92] Hans P Op de Beeck, Jennifer A Deutsch, Wim Vanduffel, Nancy G Kanwisher, and James J DiCarlo. A stable topography of selectivity for unfamiliar shape classes in monkey inferior temporal cortex. *Cerebral Cortex*, 18(7):1676–1694, 2008.

[93] Job van den Hurk, Marc Van Baelen, and Hans P Op de Beeck. Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences*, 114(22):E4501–E4510, 2017.

[94] Golijeh Golarai, Alina Liberman, and Kalanit Grill-Spector. Experience shapes the development of neural substrates of face processing in human ventral temporal cortex. *Cerebral Cortex*, 27(2): bhv314, 2015.

[95] Tina T Liu, Adrian Nestor, Mark D Vida, John A Pyles, Christina Patterson, Ying Yang, Fan Nils Yang, Erez Freud, and Marlene Behrmann. Successful reorganization of category-selective visual cortex following occipito-temporal lobectomy in childhood. *Cell reports*, 24(5):1113–1122, 2018.

[96] 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.

[97] Nooshin Abbasi, John Duncan, and Reza Rajimehr. Genetic influence is linked to cortical morphology in category-selective areas of visual cortex. *Nature Communications*, 11(1):709, 2020.

[98] Alfonso Nieto-Castañón and Evelina Fedorenko. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *Neuroimage*, 63(3):1646–1669, 2012.

[99] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[100] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[101] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

[102] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[103] 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.

[104] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[105] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

[106] Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matt Botvinick, Jane Wang, and Eric Schulz. Meta-in-context learning in large language models. *Advances in Neural Information Processing Systems*, 36:65189–65201, 2023.

[107] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

[108] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

[109] Jianlin Su. Analyzing the scale operation of attention from the perspective of entropy invariance, Dec 2021. URL https://kexue.fm/archives/8823.

[110] David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. *arXiv preprint arXiv:2202.12172*, 2022.

[111] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[112] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[113] Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, 6 (1):1–18, 2019.

[114] Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1):9383, 2024.

[115] Ken Shirakawa, Yoshihiro Nagano, Misato Tanaka, Shuntaro C Aoki, Kei Majima, Yusuke Muraki, and Yukiyasu Kamitani. Spurious reconstruction from brain activity. *arXiv preprint arXiv:2405.10078*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are reflected and addressed in the experiments section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation of the work is discussed in the discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The learning theory is described in the methods section. The model training details are discussed in the experiment setup section and in the supplemental material section. We will release the code and checkpoints before the meeting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All the data, code and checkpoints will be published before the meeting (with GitHub link included). Also, the instructions to reproduce experimental results are provided in the experimental results and supplemental material sections.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The major experimental settings including how to train and test the model is described in the experiments section. More details are provided in the supplemental material section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The experiment results are accompanied by error bars in the experiments section and supplemental material section when available.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details on computational resources are specified in the supplemental material section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The points are discussed in the discussion section and supplemental material section.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: To the best of our knowledge, the paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators or original owners of assets are properly credited and the license and terms of use are explicitly mentioned and properly respected in the main text and in the supplemental material.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release our model's code via our anonymous GitHub URL.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No new data collection or crowdsourcing was performed. Our analyses use only the public BOLD5000 and NSD dataset.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We did not collect any new human-subject data. All analyses are performed on the publicly available BOLD5000 and NSD fMRI datasets, which require no additional IRB review.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Technical Appendices and Supplementary Material

Sections

A.1 Social impacts

Our work introduces BraInCoRL, a meta-learning framework that uses fMRI-measured voxel responses and trained visual encoders to perform in-context adaptation: given a small support set of image–response pairs, the model directly estimates each voxel's response-function parameters for novel stimuli. BraInCoRL's fusion of diverse image data and neural measurements uncovers data-driven principles of cortical organization beyond both traditional hypothesis-driven experiments and computational encoding models that require thousands of samples. Moreover, its alignment of neural responses with natural language prompts enables the generation of new hypotheses about semantic representation in visual cortex. As such, BraInCoRL may accelerate early diagnosis and monitoring of visual or neurological disorders via rapid, subject-specific cortical mapping; guide more efficient experimental design through optimized stimulus selection; deepen our understanding of semantic coding and inter-subject variability; and, when integrated with generative image models, open avenues for brain-guided stimulus synthesis and personalized neuroprosthetic and brain–computer interface development. While BraInCoRL offers significant potential for neuroscience and clinical applications, its reliance on fMRI datasets and computational infrastructure may limit accessibility for under-resourced research groups and raise privacy concerns if applied to sensitive neural data.

A.2 Implementation details

Network architecture. Our BraInCoRL model architecture comprises three main components:

1. **Input projection.** An input projection MLP layer is applied to each token individually, which maps the stimulus semantics and voxel activation into an embedding. In detail, we concatenate each image embedding with its scalar neural response and pass it through a single-layer MLP to align the two modalities into a unified internal feature space.

2. **Transformer encoder.** A stack of 20 self-attention layers integrates information across all support examples (plus learnable tokens), capturing contextual relationships and the relative importance of each in-context sample. We adopt best practices and utilize SwiGLU activation paired with pre-normalization in the attention block. We utilize 10 heads in each multi-head self attention.

3. **Weight prediction.** The [CLS] token from the final layer goes through an MLP to output a hyperweight which is used to parameterize the final encoder. In detail, the aggregated representation is fed through another single-layer MLP that outputs a weight vector, which is then used to linearly combine unknown-image embeddings to produce the final neural response predictions.

The CLIP-based variant (encoding dimension $E = 512$) contains approximately 97.2 M parameters; DINO ($E = 768$) and SIGLIP ($E = 1152$) variants comprise roughly 112 M and 130 M parameters, respectively.

Model training. Training is implemented in PyTorch on eight NVIDIA RTX 6000 Ada GPUs (48 GB each). We optimize using AdamW (decoupled weight decay 1×10^{-4}) with an initial learning rate of 1×10^{-3}, which decays to 1×10^{-5} via a ReduceLROnPlateau scheduler (factor 0.1, patience 5 epochs, cooldown 2 epochs, threshold 1e-4). Mini-batches randomly sample an in-context support set of 100 images in the first pretraining stage, and randomly sample between 30 and 500 in-context support images in the second context extension stage and the third finetuning stages. Each training stage runs for up to 100 epochs with early stopping based on validation loss (patience: 5 epochs). The training batch size is 80. We allocate 20% of the test set data for validation.

Computational cost. With an in-context support set of 100 images, our model predicts responses for $\sim$ 20,000 voxels in the higher visual cortex in under 20 seconds on a single RTX 6000 Ada GPU.

A.3 Text prompts for category-selective regions

We define a set of natural language prompts for each semantic category. For every image–prompt pair, we use the CLIP text encoder to generate text encodings. The natural language prompts for each category are listed below:

Faces [A photo of a person's face, A portrait photo of a face, A face facing the camera, A photo of a face, A photo of a human face, A photo of faces, People looking at the camera, A portrait of a person, A portrait photo]

Bodies [A photo of a torso, A photo of limbs, A photo of bodies, A photo of a person, A photo of people, A photo of a body, A person's arms, A person's legs, A photo of hands]

Places [A photo of a bedroom, A photo of an office, A photo of a hallway, A photo of a doorway, A photo of interior design, A photo of a building, A photo of a house, A photo of nature, A photo of a landscape]

Food [A photo of food, A photo of cuisine, A photo of fruit, A photo of foodstuffs, A photo of a meal, A photo of bread, A photo of rice, A photo of a snack, A photo of pastries]

Words [A photo of words, A photo of glyphs, A photo of a glyph, A photo of text, A photo of numbers, A photo of a letter, A photo of letters, A photo of writing, A photo of text on an object]

A.4 A more detailed description of logit scaling

The motivating factor underlying logit scaling is our desire to have our in-context learned encoder perform well regardless of the number of stimuli given to the model, and effectively generalize to context sizes beyond those seen during training. For example, while we may only train with between 30 and 500 images, a third-party experimenter may want to use fewer than 30 images or more than 500 images to condition the model. Across all cases, we want the model to succeed.

This logit scaling method was first proposed in [1, 2], and later adapted in Qwen LLM (logn-scaling) [3] and Llama 4 LLM (temperature scaling) [4]. We will briefly summarize the high-level math, which we take from [1] with our commentary:

Let the attention value in self-attention for a particular query token i to value token j to be

$$a_{i,j} = \frac{e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j}}{\sum_{j=1}^n e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j}}$$

Then the entropy is defined as

$$\mathcal{H}_i = -\sum_{j=1}^n a_{i,j} \log a_{i,j}$$

Substituting the expression for $a_{i,j}$ we have the entropy as

$$\mathcal{H}_i = -\sum_{j=1}^n a_{i,j} \log \left(\frac{e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j}}{\sum_{j=1}^n e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j}} \right)$$

Since $\sum_{j=1}^n a_{i,j} = 1$,

we can express the formula as

$$\mathcal{H}_i = \log \sum_{j=1}^n e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j} - \frac{\sum_{j=1}^n e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j} (\lambda \mathbf{q}_i \cdot \mathbf{k}_j)}{\sum_{j=1}^n e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j}}$$

If the first term is expressed as an expectation over j, we have

$$\sum_{j=1}^n e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j} = n \times \frac{1}{n} \sum_{j=1}^n e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j} \approx n \mathbb{E}_j[e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j}]$$

Which leads to the following approximation of entropy

$$\mathcal{H}_i \approx \log n + \log \mathbb{E}_j[e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j}] - \frac{\lambda \mathbb{E}_j[e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j} (\mathbf{q}_i \cdot \mathbf{k}_j)]}{\mathbb{E}_j[e^{\lambda \mathbf{q}_i \cdot \mathbf{k}_j}]}$$

If the vectors are assumed to be the output of a layernorm layer with length $\sqrt{d}$, the expectation can be converted to one over the angles between vectors

$$\mathcal{H}_i \approx \log n + \log \mathbb{E}_\theta[e^{\lambda d \cos \theta}] - \frac{\lambda d \mathbb{E}_\theta[e^{\lambda d \cos \theta} \cos \theta]}{\mathbb{E}_\theta[e^{\lambda d \cos \theta}]}$$

Since most randomly distributed vectors in higher dimensions are orthogonal, we derive a term which can roughly be expressed as

$$\mathcal{H}_i \approx \log n + C$$

where C does not depend on the number of tokens n.

This leads to an approximate scaling factor for the logits of $\log n$ to keep the entropy invariant to context length.

Therefore, we change the standard formulation of attention values by applying a scaling factor of $\log n$ to the QK^T term, as shown in Equation (4) of our paper.

Note that in the above explanation we adopt the notation from [1].

A.5 Cortex prediction explained variance for different image encoding backbones and subjects

In this section, we compare three methods across multiple subjects (S1-S8) and embedding backbones (CLIP, DINO, SigLIP): the fully trained reference model fit to converge on each subject's full 9,000-image training set; our BraInCoRL approach, which adapts to a new subject with only 100 in-context images; and a within-subject ridge regression baseline trained on the same 100 images with the BraInCoRL approach. In every case, BraInCoRL outperforms ridge regression and achieves accuracy similar to that of the fully trained model.
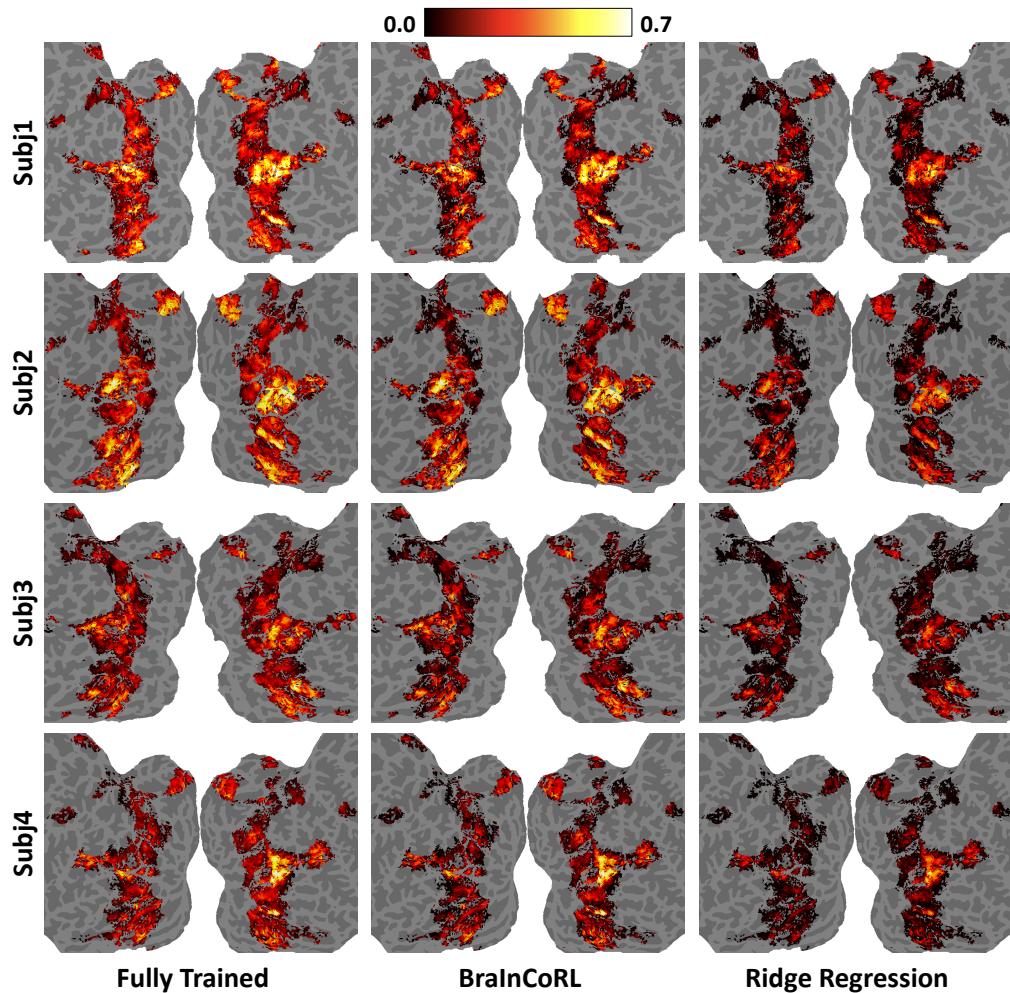
Figure S.1: **Higher visual cortex explained variance of CLIP backbone.** From left to right, we show the explained variance from the full model trained on 9000 images for a subject, BraInCoRL with just 100 in-context images from the new subject, and within-subject ridge regression with 100 images using CLIP backbone for subject 1,2,3,4.
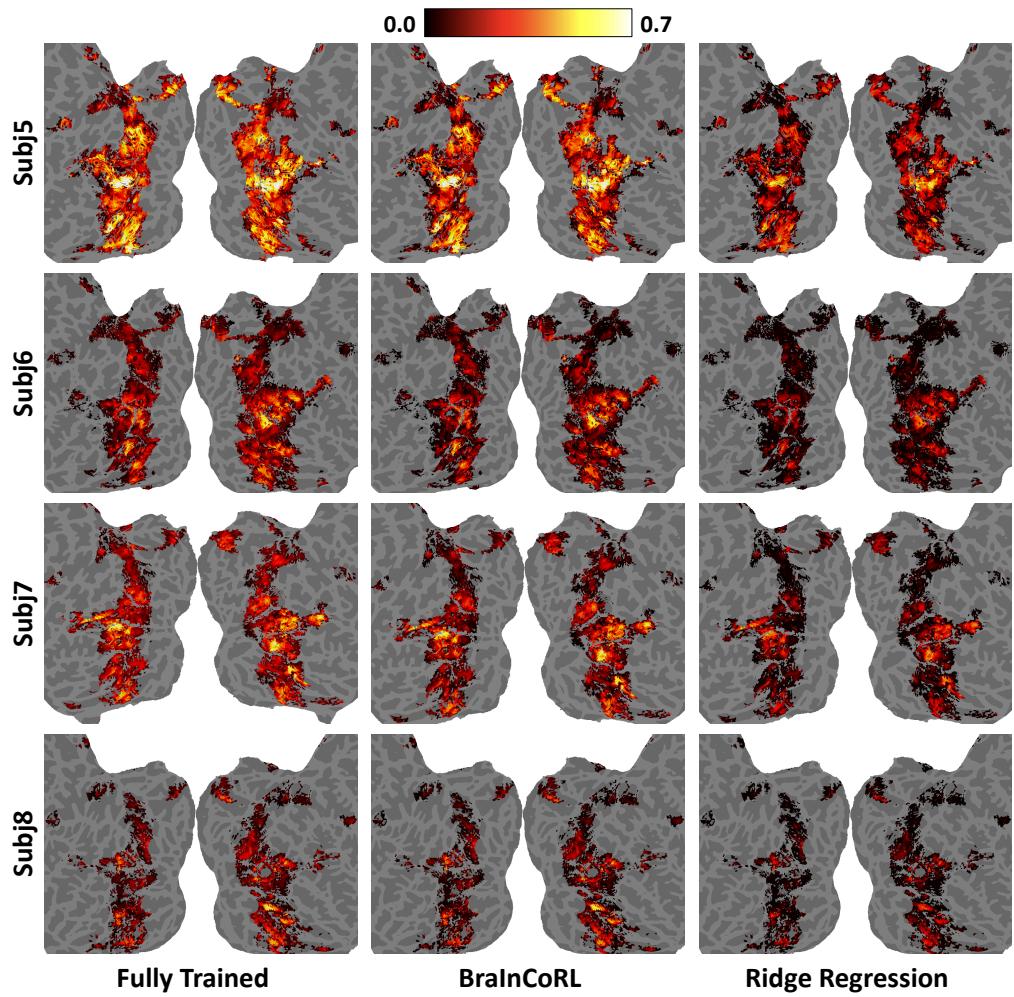
Figure S.2: **Higher visual cortex explained variance of CLIP backbone.** From left to right, we show the explained variance from the full model trained on 9000 images for a subject, BraInCoRL with just 100 in-context images from the new subject, and within-subject ridge regression with 100 images using CLIP backbone for subject 5, 6, 7, 8.
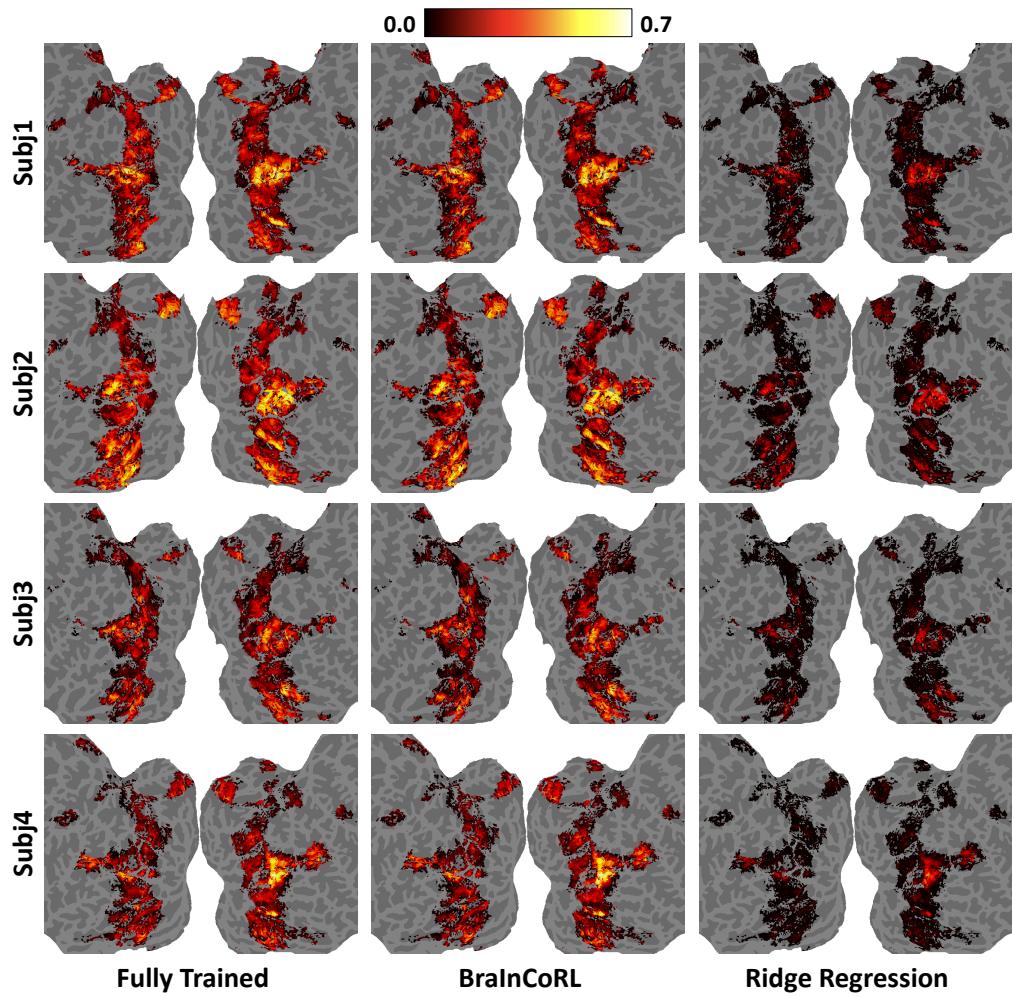
Figure S.3: **Higher visual cortex explained variance of DINO backbone.** From left to right, we show the explained variance from the full model trained on 9000 images for a subject, BraInCoRL with just 100 in-context images from the new subject, and within-subject ridge regression with 100 images using DINO backbone for subject 1, 2, 3, 4.
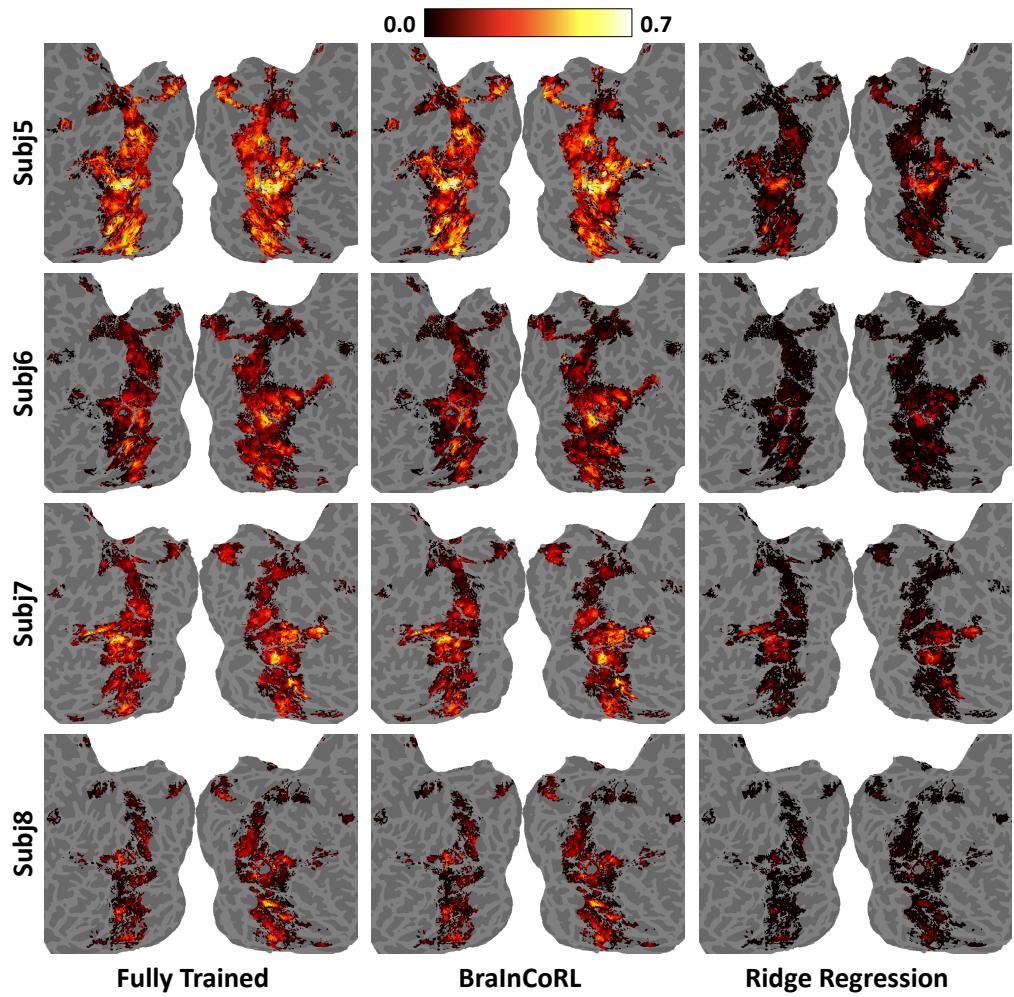
Figure S.4: **Higher visual cortex explained variance of DINO backbone.** From left to right, we show the explained variance from the full model trained on 9000 images for a subject, BraInCoRL with just 100 in-context images from the new subject, and within-subject ridge regression with 100 images using DINO backbone for subject 5, 6, 7, 8.
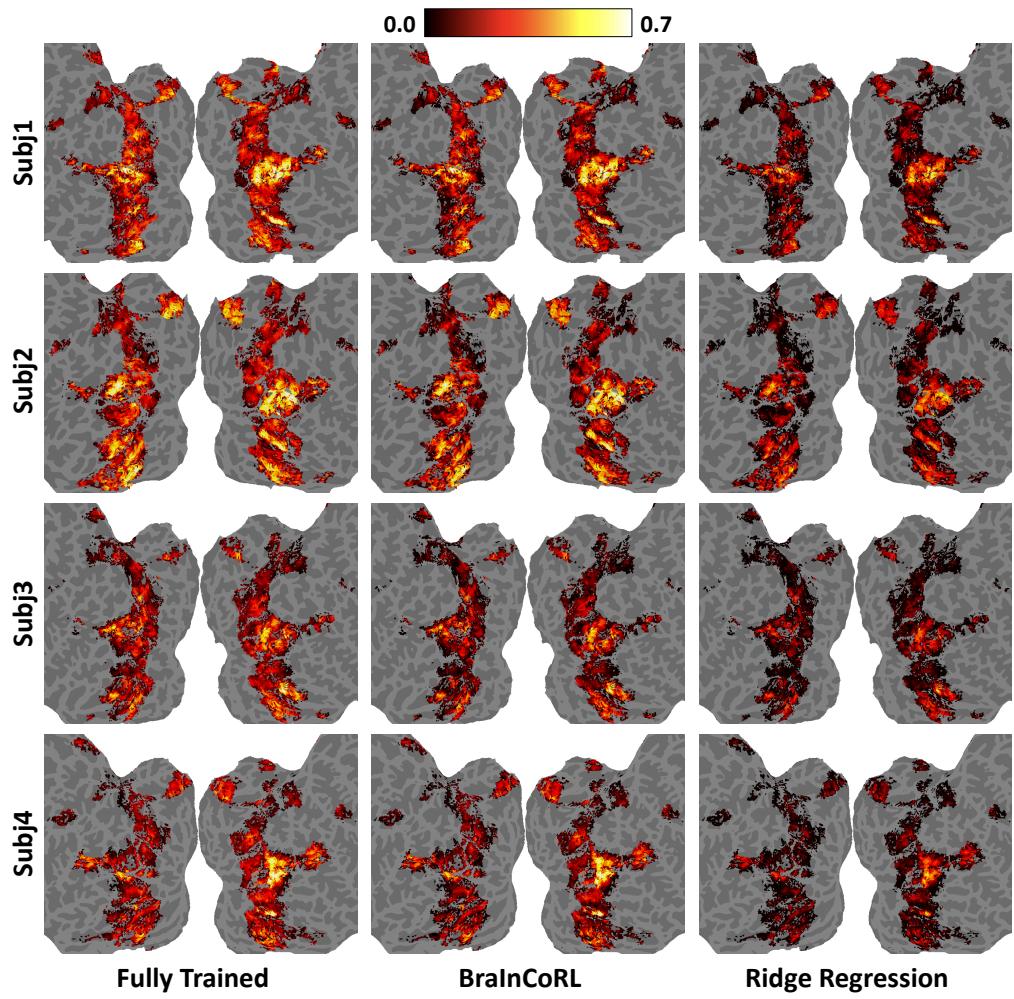
Figure S.5: **Higher visual cortex explained variance of SigLIP backbone.** From left to right, we show the explained variance from the full model trained on 9000 images for a subject, BraInCoRL with just 100 in-context images from the new subject, and within-subject ridge regression with 100 images using SigLIP backbone for subject 1, 2, 3, 4.
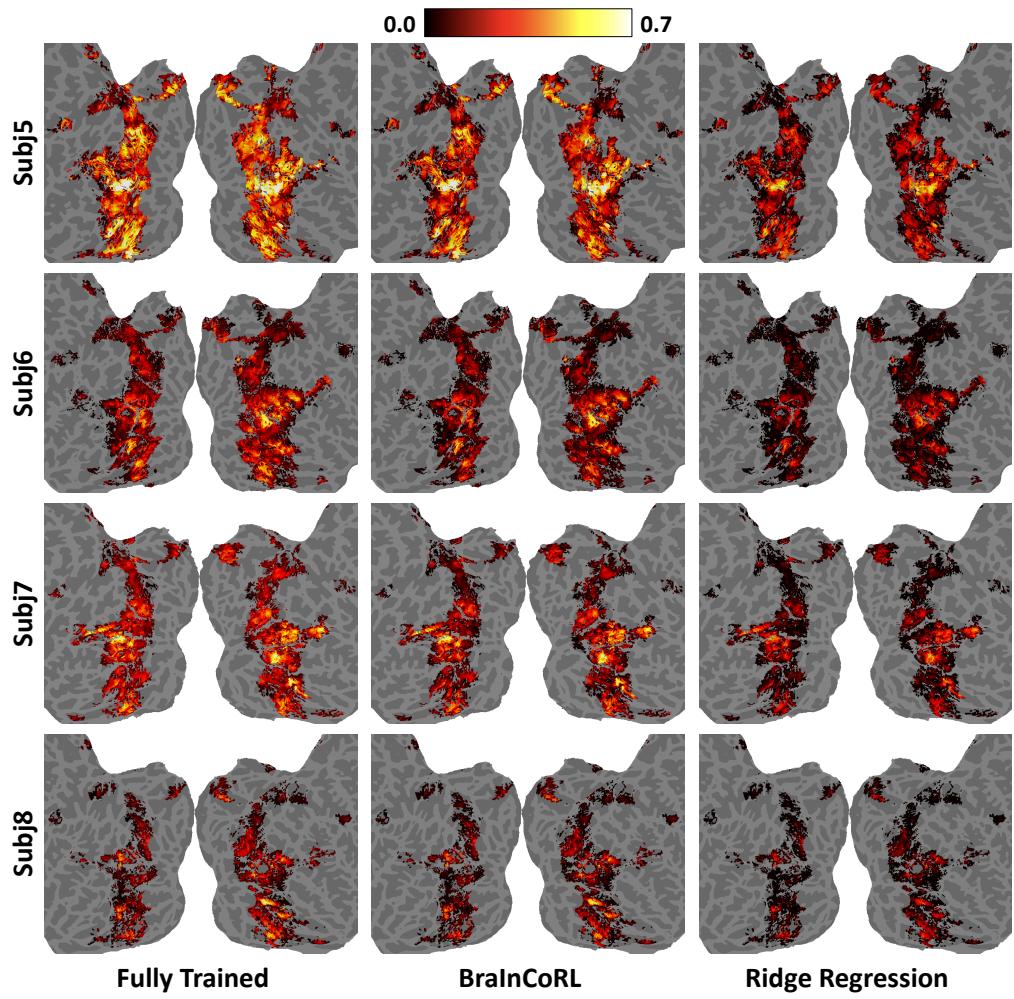
Figure S.6: **Higher visual cortex explained variance of SigLIP backbone.** From left to right, we show the explained variance from the full model trained on 9000 images for a subject, BraInCoRL with just 100 in-context images from the new subject, and within-subject ridge regression with 100 images using SigLIP backbone for subject 5, 6, 7, 8.

A.6 Voxelwise performance across five category-selective regions for different image encoding backbones and subjects

In this section, we report voxel-wise explained variance in five category-selective regions (faces, places, bodies, words, and food) with CLIP, DINO and SigLIP backbone for subjects S1-S8. We compare our in-context model (BraInCoRL) against the fully trained reference model fit to converge on each subject's full 9,000-image training set, within-subject ridge regression baselines trained on 100 and 300 images, and the FsAverage map. BraInCoRL outperforms the ridge baselines and closely approaches the performance of the fully trained model.

Table S.1: **Voxel-wise explained variance with the CLIP backbone for Subjects 1 and 2.** We report performance for our in-context model (BraInCoRL), the fully trained reference ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**
Fully Trained	0.19	0.16	0.20	0.27	0.28	0.24	0.11	0.11	0.16	0.17	0.18	0.19
Ridge-100	0.10	0.07	0.08	0.14	0.16	0.12	0.02	0.03	0.05	0.07	0.07	0.08
Ridge-300	0.13	0.10	0.13	0.20	0.22	0.16	0.06	0.06	0.10	0.11	0.11	0.12
FsAverage map	0.13	0.06	0.11	0.19	0.09	0.08	0.06	0.03	0.14	0.18	0.08	0.06
BraInCoRL-100	**0.16**	**0.13**	**0.16**	**0.23**	**0.25**	**0.21**	**0.07**	**0.08**	**0.12**	**0.13**	**0.13**	**0.15**

Table S.2: **Voxel-wise explained variance with the CLIP backbone for Subjects 3 and 4.** We report performance for our in-context model (BraInCoRL), the fully trained reference ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S3	**S4**	**S3**	**S4**	**S3**	**S4**	**S3**	**S4**	**S3**	**S4**	**S3**	**S4**
Fully Trained	0.16	0.14	0.16	0.16	0.17	0.17	0.09	0.07	0.10	0.12	0.13	0.14
Ridge-100	0.07	0.05	0.08	0.05	0.08	0.08	0.02	0.01	0.03	0.04	0.05	0.05
Ridge-300	0.10	0.09	0.11	0.10	0.11	0.12	0.04	0.04	0.05	0.07	0.08	0.09
FsAverage map	0.10	0.03	0.14	0.05	0.11	0.06	0.07	0.03	0.10	0.07	0.10	0.04
BraInCoRL-100	**0.12**	**0.10**	**0.13**	**0.13**	**0.14**	**0.13**	**0.05**	**0.04**	**0.07**	**0.08**	**0.10**	**0.10**

Table S.3: **Voxel-wise explained variance with the CLIP backbone for Subjects 5 and 6.** We report performance for our in-context model (BraInCoRL), the fully trained reference ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S5	**S6**	**S5**	**S6**	**S5**	**S6**	**S5**	**S6**	**S5**	**S6**	**S5**	**S6**
Fully Trained	0.24	0.17	0.32	0.13	0.26	0.18	0.17	0.09	0.24	0.09	0.23	0.11
Ridge-100	0.11	0.07	0.16	0.03	0.13	0.09	0.06	0.01	0.11	0.02	0.10	0.03
Ridge-300	0.16	0.11	0.24	0.07	0.19	0.13	0.10	0.04	0.16	0.04	0.15	0.06
FsAverage map	0.07	0.05	0.11	0.08	0.06	0.04	0.05	0.06	0.08	0.04	0.07	0.05
BraInCoRL-100	**0.20**	**0.14**	**0.29**	**0.10**	**0.23**	**0.15**	**0.12**	**0.05**	**0.19**	**0.05**	**0.19**	**0.08**

Table S.4: **Voxel-wise explained variance with the CLIP backbone for Subjects 7 and 8.** We report performance for our in-context model (BraInCoRL), the fully trained oracle ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S7	S8	S7	S8	S7	S8	S7	S8	S7	S8	S7	S8
Fully Trained	0.14	0.08	0.18	0.10	0.21	0.09	0.12	0.04	0.12	0.06	0.16	0.09
Ridge-100	0.06	0.03	0.08	0.04	0.11	0.04	0.03	0.00	0.03	0.02	0.06	0.03
Ridge-300	0.09	0.05	0.12	0.06	0.15	0.05	0.06	0.01	0.06	0.03	0.09	0.05
FsAverage map	0.12	0.04	0.17	0.04	0.10	0.03	0.09	0.03	0.19	0.02	0.09	0.03
BraInCoRL-100	**0.11**	**0.07**	**0.15**	**0.08**	**0.18**	**0.07**	**0.08**	**0.02**	**0.08**	**0.04**	**0.12**	**0.07**

Table S.5: **Voxel-wise explained variance with the DINO backbone for Subjects 1 and 2.** We report performance for our in-context model (BraInCoRL), the fully trained oracle ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fully Trained	0.15	0.13	0.16	0.22	0.24	0.20	0.08	0.08	0.13	0.14	0.14	0.15
Ridge-100	0.03	0.02	0.03	0.05	0.07	0.05	0.00	0.01	0.02	0.03	0.02	0.03
Ridge-300	0.07	0.05	0.07	0.10	0.14	0.10	0.03	0.03	0.05	0.05	0.06	0.06
FsAverage map	0.13	0.06	0.11	0.19	0.09	0.08	0.06	0.03	0.14	0.18	0.08	0.06
BraInCoRL-100	**0.14**	**0.12**	**0.15**	**0.21**	**0.23**	**0.18**	**0.07**	**0.07**	**0.11**	**0.12**	**0.12**	**0.14**

Table S.6: **Voxel-wise explained variance with the DINO backbone for Subjects 3 and 4.** We report performance for our in-context model (BraInCoRL), the fully trained oracle ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S3	S4	S3	S4	S3	S4	S3	S4	S3	S4	S3	S4
Fully Trained	0.13	0.11	0.13	0.12	0.13	0.14	0.06	0.05	0.08	0.09	0.10	0.11
Ridge-100	0.02	0.00	0.03	0.01	0.03	0.02	-0.01	-0.02	0.00	-0.02	0.01	0.01
Ridge-300	0.05	0.04	0.06	0.04	0.06	0.07	0.02	0.01	0.03	0.03	0.04	0.04
FsAverage map	0.10	0.03	0.14	0.05	0.11	0.06	0.07	0.03	0.10	0.07	0.09	0.04
BraInCoRL-100	**0.11**	**0.10**	**0.13**	**0.12**	**0.13**	**0.14**	**0.05**	**0.05**	**0.06**	**0.08**	**0.09**	**0.10**

Table S.7: **Voxel-wise explained variance with the DINO backbone for Subjects 5 and 6.** We report performance for our in-context model (BraInCoRL), the fully trained reference ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S5	S6	S5	S6	S5	S6	S5	S6	S5	S6	S5	S6
Fully Trained	0.19	0.14	0.27	0.11	0.22	0.15	0.13	0.06	0.20	0.06	0.19	0.08
Ridge-100	0.04	0.02	0.05	0.01	0.07	0.03	0.01	-0.01	0.03	-0.01	0.04	0.01
Ridge-300	0.08	0.06	0.13	0.04	0.12	0.07	0.04	0.01	0.08	0.02	0.09	0.03
FsAverage map	0.07	0.05	0.11	0.08	0.06	0.04	0.05	0.06	0.08	0.04	0.07	0.05
BraInCoRL-100	**0.18**	**0.12**	**0.26**	**0.10**	**0.21**	**0.14**	**0.10**	**0.04**	**0.18**	**0.04**	**0.17**	**0.07**

Table S.8: **Voxel-wise explained variance with the DINO backbone for Subjects 7 and 8.** We report performance for our in-context model (BraInCoRL), the fully trained oracle ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S7	S8	S7	S8	S7	S8	S7	S8	S7	S8	S7	S8
Fully Trained	0.10	0.06	0.15	0.08	0.18	0.07	0.08	0.02	0.09	0.03	0.12	0.07
Ridge-100	0.02	-0.00	0.02	0.00	0.06	0.00	0.01	-0.01	0.00	-0.00	0.02	-0.00
Ridge-300	0.05	0.02	0.06	0.03	0.10	0.03	0.03	0.00	0.02	0.01	0.05	0.02
FsAverage map	0.12	0.04	0.17	0.04	0.10	0.03	0.09	0.03	0.19	0.02	0.09	0.03
BraInCoRL-100	**0.10**	**0.06**	**0.14**	**0.07**	**0.17**	**0.07**	**0.07**	**0.02**	**0.07**	**0.04**	**0.10**	**0.06**

Table S.9: **Voxel-wise explained variance with the SigLIP backbone for Subjects 1 and 2.** We report performance for our in-context model (BraInCoRL), the fully trained oracle ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fully Trained	0.19	0.17	0.21	0.27	0.30	0.25	0.12	0.11	0.17	0.18	0.19	0.20
Ridge-100	0.10	0.07	0.09	0.14	0.18	0.12	0.03	0.04	0.06	0.07	0.08	0.08
Ridge-300	0.14	0.11	0.14	0.20	0.23	0.17	0.07	0.06	0.11	0.12	0.12	0.13
FsAverage map	0.13	0.06	0.11	0.19	0.09	0.08	0.06	0.03	0.14	0.18	0.08	0.06
BraInCoRL-100	**0.17**	**0.13**	**0.18**	**0.24**	**0.27**	**0.21**	**0.09**	**0.08**	**0.13**	**0.14**	**0.15**	**0.16**

Table S.10: **Voxel-wise explained variance with the SigLIP backbone for Subjects 3 and 4.** We report performance for our in-context model (BraInCoRL), the fully trained oracle ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S3	S4	S3	S4	S3	S4	S3	S4	S3	S4	S3	S4
Fully Trained	0.17	0.14	0.17	0.17	0.18	0.18	0.10	0.08	0.11	0.13	0.14	0.15
Ridge-100	0.07	0.04	0.08	0.06	0.09	0.07	0.02	0.01	0.03	0.04	0.05	0.05
Ridge-300	0.11	0.08	0.11	0.10	0.12	0.12	0.05	0.03	0.06	0.07	0.08	0.09
FsAverage map	0.10	0.03	0.14	0.05	0.11	0.06	0.07	0.03	0.10	0.07	0.10	0.04
BraInCoRL-100	**0.12**	**0.11**	**0.13**	**0.14**	**0.14**	**0.15**	**0.05**	**0.05**	**0.06**	**0.09**	**0.10**	**0.12**

Table S.11: **Voxel-wise explained variance with the SigLIP backbone for Subjects 5 and 6.** We report performance for our in-context model (BraInCoRL), the fully trained reference ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S5	S6	S5	S6	S5	S6	S5	S6	S5	S6	S5	S6
Fully Trained	0.24	0.18	0.33	0.15	0.28	0.19	0.18	0.10	0.26	0.10	0.24	0.12
Ridge-100	0.11	0.07	0.16	0.04	0.14	0.09	0.06	0.01	0.11	0.02	0.11	0.04
Ridge-300	0.17	0.12	0.24	0.08	0.20	0.14	0.10	0.04	0.17	0.05	0.16	0.07
FsAverage map	0.07	0.05	0.11	0.08	0.06	0.04	0.05	0.06	0.08	0.04	0.07	0.05
BraInCoRL-100	**0.20**	**0.14**	**0.28**	**0.11**	**0.23**	**0.16**	**0.12**	**0.05**	**0.19**	**0.05**	**0.19**	**0.09**

Table S.12: **Voxel-wise explained variance with the SigLIP backbone for Subjects 7 and 8.** We report performance for our in-context model (BraInCoRL), the fully trained oracle ("Fully Trained"), within-subject ridge regression baselines (100, 300), and the FsAverage map across five category-selective regions (faces, places, bodies, words, food).

	Faces		Places		Bodies		Words		Food		Mean	
	S7	**S8**	**S7**	**S8**	**S7**	**S8**	**S7**	**S8**	**S7**	**S8**	**S7**	**S8**
Fully Trained	0.14	0.09	0.15	0.10	0.22	0.09	0.13	0.04	0.13	0.06	0.16	0.09
Ridge-100	0.06	0.02	0.08	0.04	0.11	0.03	0.03	0.00	0.03	0.02	0.06	0.03
Ridge-300	0.10	0.05	0.12	0.07	0.15	0.05	0.06	0.02	0.06	0.03	0.09	0.05
FsAverage map	0.12	0.04	0.17	0.04	0.10	0.03	0.09	0.03	0.19	0.02	0.09	0.03
BraInCoRL-100	**0.11**	**0.07**	**0.15**	**0.08**	**0.19**	**0.08**	**0.08**	**0.02**	**0.09**	**0.04**	**0.12**	**0.07**

A.7 Voxelwise explained variance across varying support set sizes for more subjects, backbones and for pretrain-only models

In this section, we investigate how the size of the in-context support set affects voxelwise predictive performance. We evaluate three image-encoding backbones (CLIP, DINO, SigLIP) on eight subjects (S1–S8) by comparing the performance of BraInCoRL with the pretrain-only BraInCoRL (i.e. same architecture but only pretrained), the within-subject ridge-regression baseline, and the fully trained reference model fit to converge on each subject's full 9,000-image training set.
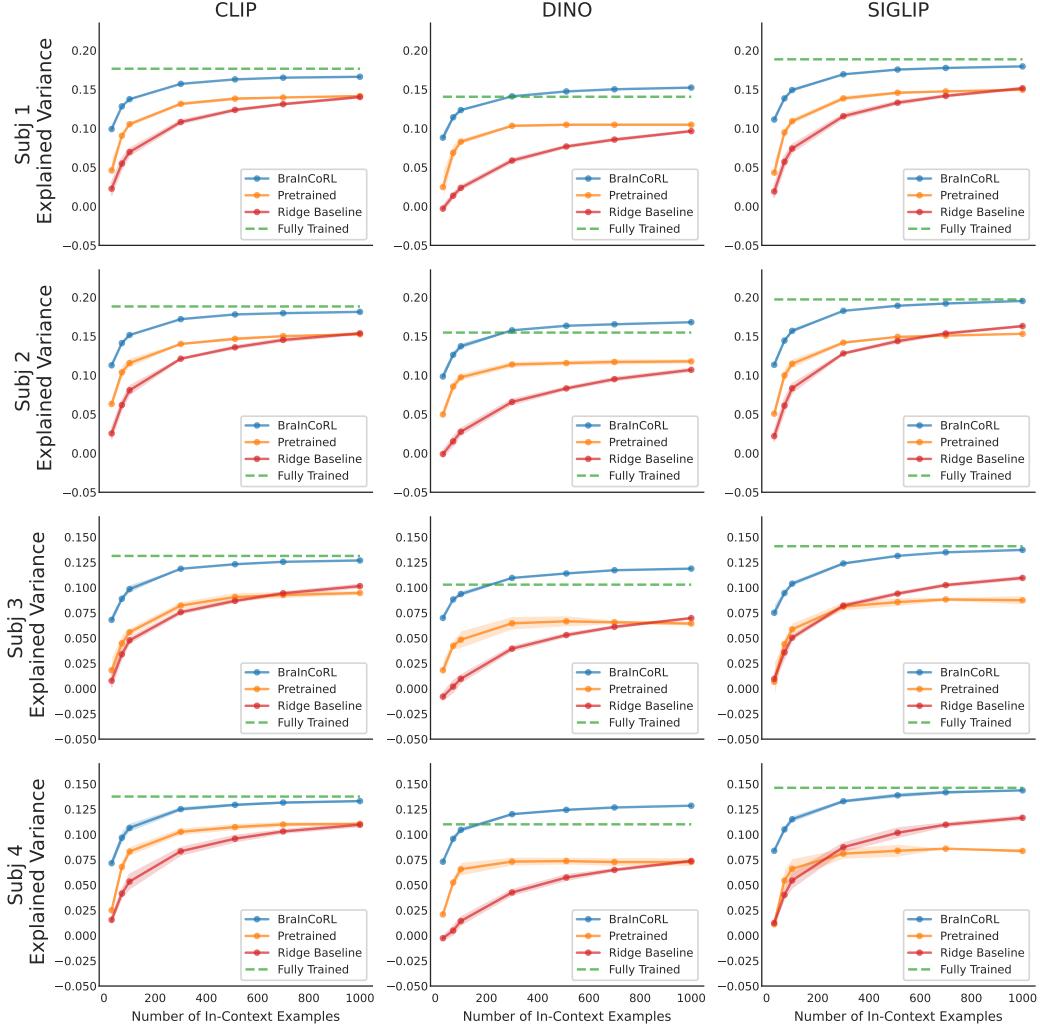
Figure S.7: **Voxelwise explained variance as a function of in-context support set size.** Voxelwise explained variance is visualized for subjects 1, 2, 3, 4 for each backbone (CLIP, DINO, SigLIP). we plot results for the BraInCoRL, along with the pretrain-only BraInCoRL (i.e. same architecture but only pretrained), the within-subject ridge-regression baseline, and the fully trained reference model using all 9,000 images.
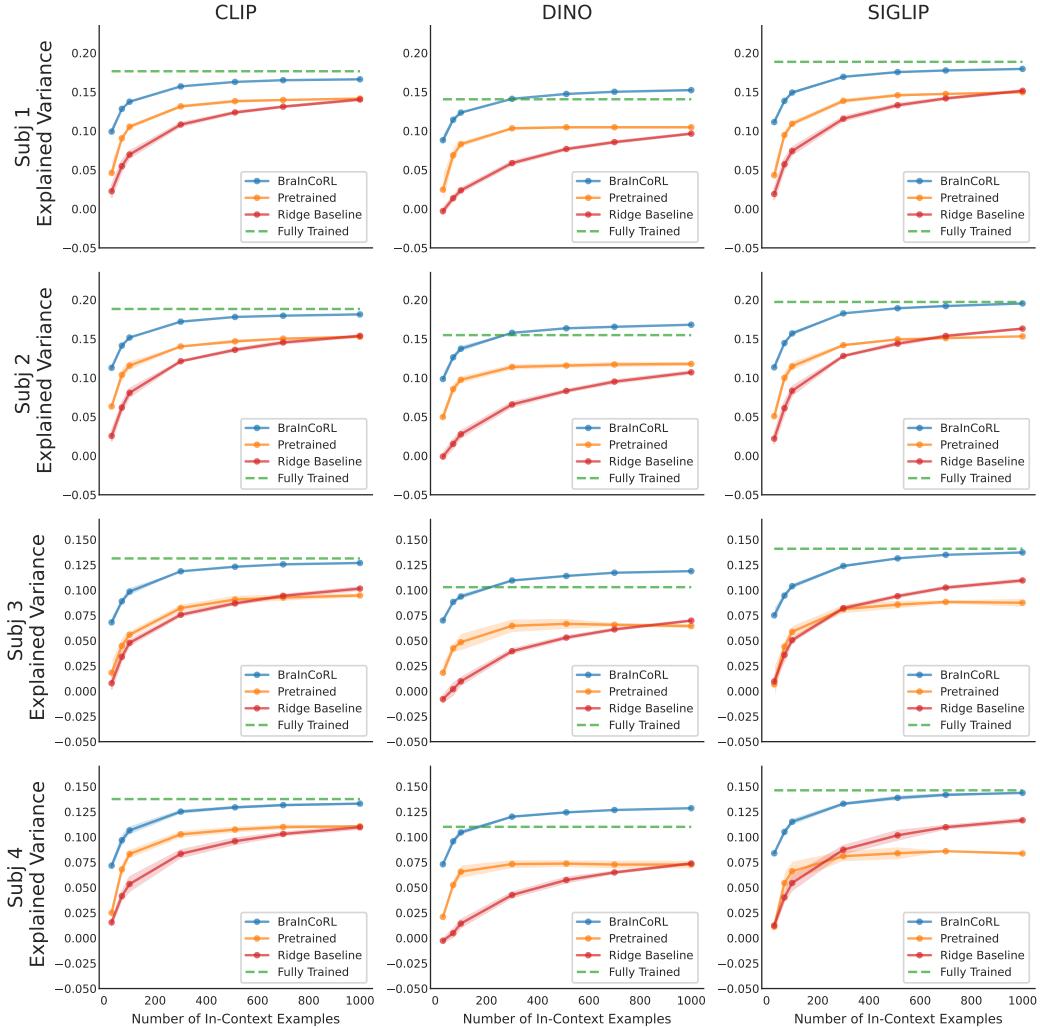
Figure S.8: **Voxelwise explained variance as a function of in-context support set size.** Voxelwise explained variance is visualized for subjects 5, 6, 7, 8 for each backbone (CLIP, DINO, SigLIP). we plot results for the BraInCoRL, along with the pretrain-only BraInCoRL (i.e. same architecture but only pretrained), the within-subject ridge-regression baseline, and the fully trained reference model using all 9,000 images.

A.8 Impact of holding out the test subject's unique images during meta-training evaluated on more backbones

In this section, we further conduct ablations by evaluating a BraInCoRL model trained without holding out the test subject's support images ("BraInCoRL no HO") and a BraInCoRL model with only pretraining, on DINO and SigLIP backbones. The result indicates that fine-tuning on real neural data enhances performance, and that BraInCoRL is able to generalize effectively to entirely unseen images without having encountered them during training.
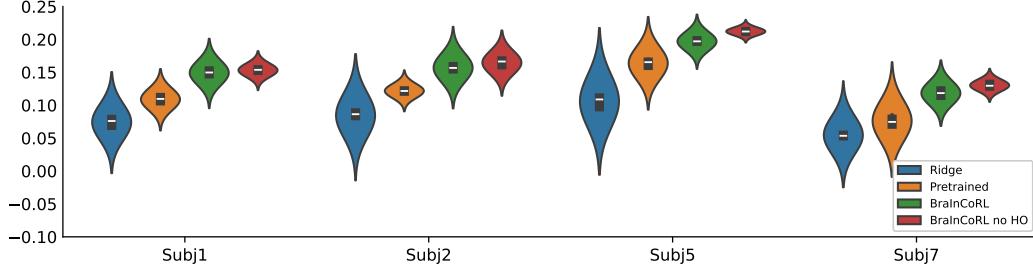
Figure S.9: **Distributions of voxelwise explained variance for subjects 1, 2, 5 and 7 using DINO encoding.** Results confirm that finetuning with real neural data boosts performance and that BraInCoRL can generalize well to previously unseen images without requiring them during training.
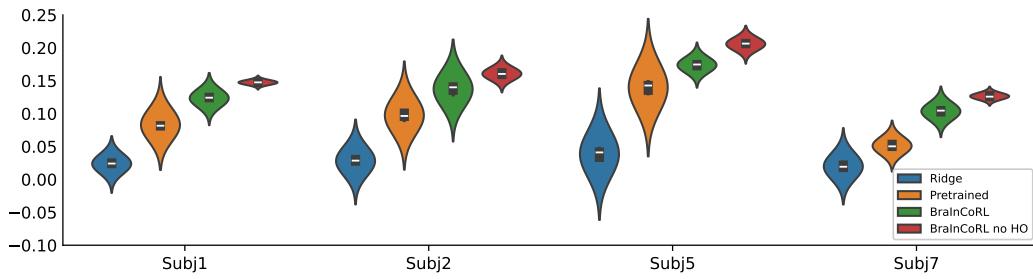
Figure S.10: **Distributions of voxelwise explained variance for subjects 1, 2, 5 and 7 using SigLIP encoding.** Results confirm that finetuning with real neural data boosts performance and that BraInCoRL can generalize well to previously unseen images without requiring them during training.

A.9 Correlation of each backbone's predictions with fully trained activation predictions

Using various image-encoding backbones, we plot how each subject's BraInCoRL predicted explained variance correlates with the fully trained model's explained variance (fully trained model refers to the fully trained reference model fit to converge on each subject's full 9,000-image training set). Across every backbone and all subjects, this correlation remains uniformly high.
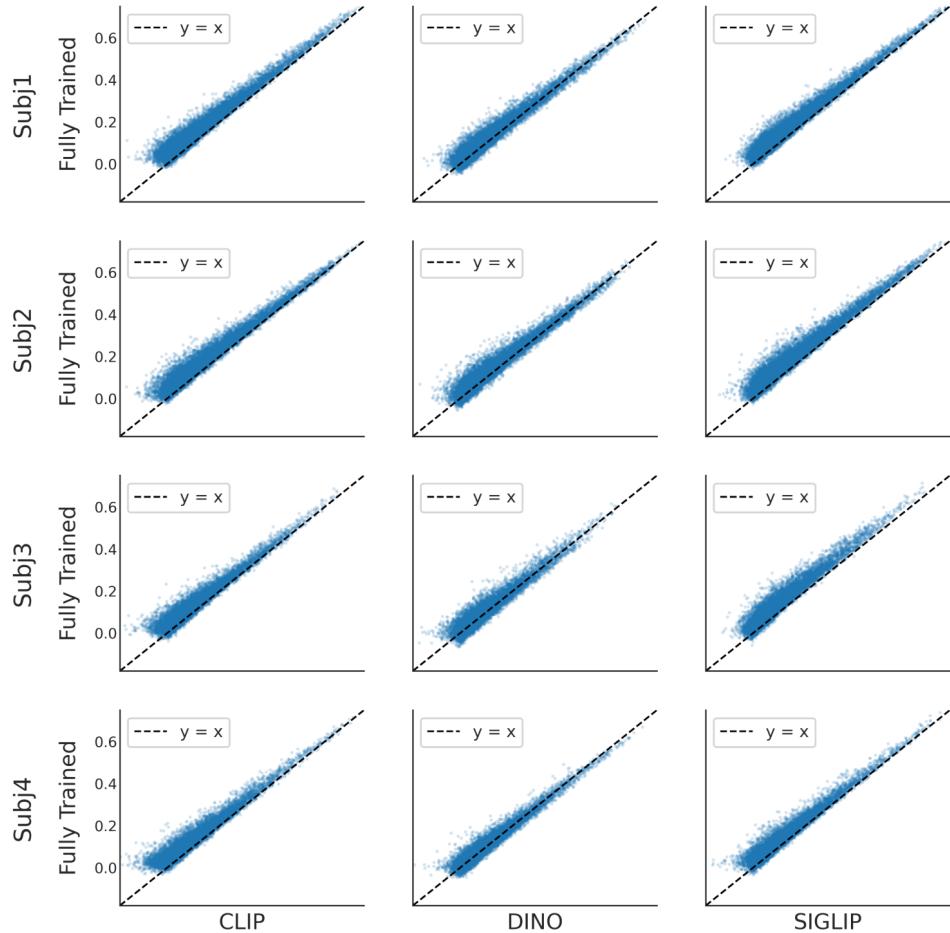
Figure S.11: **Voxelwise explained-variance correlation across backbones for subject 1, 2, 3, 4.** Each panel shows a scatter of the BraInCoRL's explained-variance predictions (100 in-context examples) versus the fully trained reference model's explained variance for each voxel. Rows correspond to subjects (1, 2, 3, 4) and columns to image-encoding backbones (CLIP, DINO, SigLIP). The dashed line marks $y = x$.
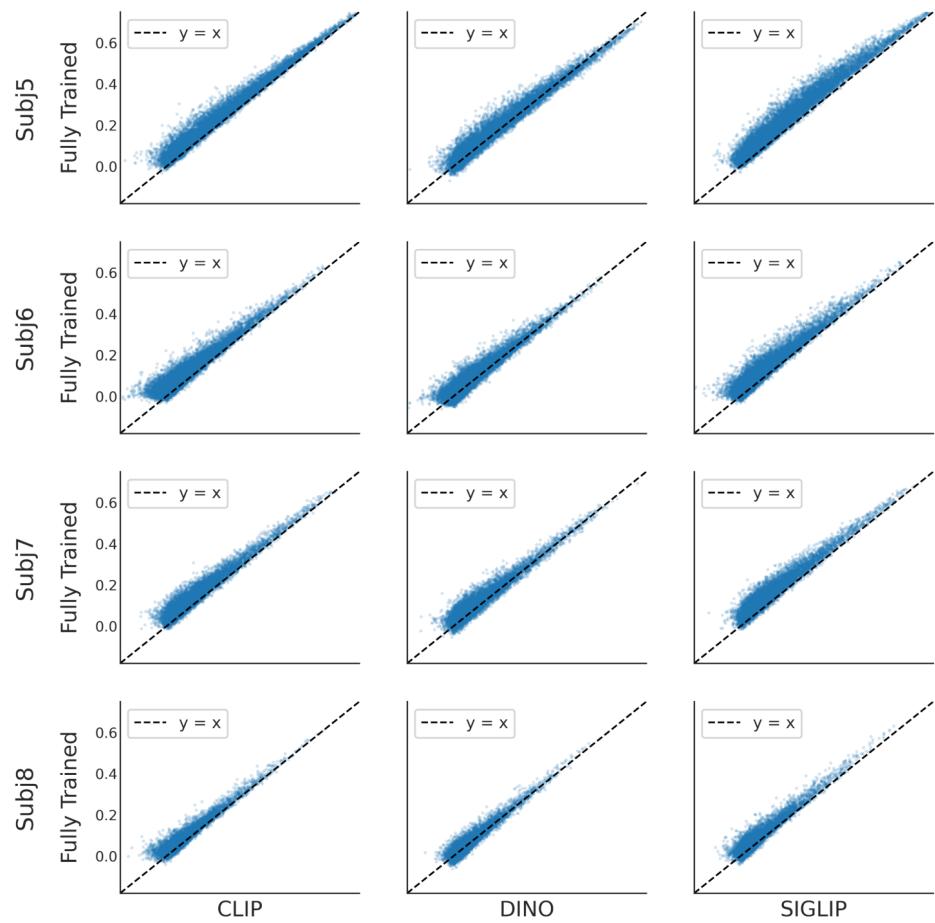
Figure S.12: **Voxelwise explained-variance correlation across backbones for subject 5, 6, 7, 8.** Each panel shows a scatter of the BraInCoRL's explained-variance predictions (100 in-context examples) versus the fully trained reference model's explained variance for each voxel. Rows correspond to subjects (5, 6, 7, 8) and columns to image-encoding backbones (CLIP, DINO, SigLIP). The dashed line marks $y = x$.

A.10 Voxelwise explained-variance evaluation in BOLD5000 for more subjects and different backbones

In this section, we analyze how varying the number of in-context examples influences voxel-level prediction performance on the BOLD5000 dataset. For subject S2 and S3, we plot the mean Pearson's r between model-predicted and actual BOLD responses as a function of support-set size, comparing our BraInCoRL model against a ridge regression baseline. Results are averaged over five cross-validation folds. Across all three image-encoding backbones (CLIP, DINO, and SIGLIP), BraInCoRL consistently outperforms ridge regression.
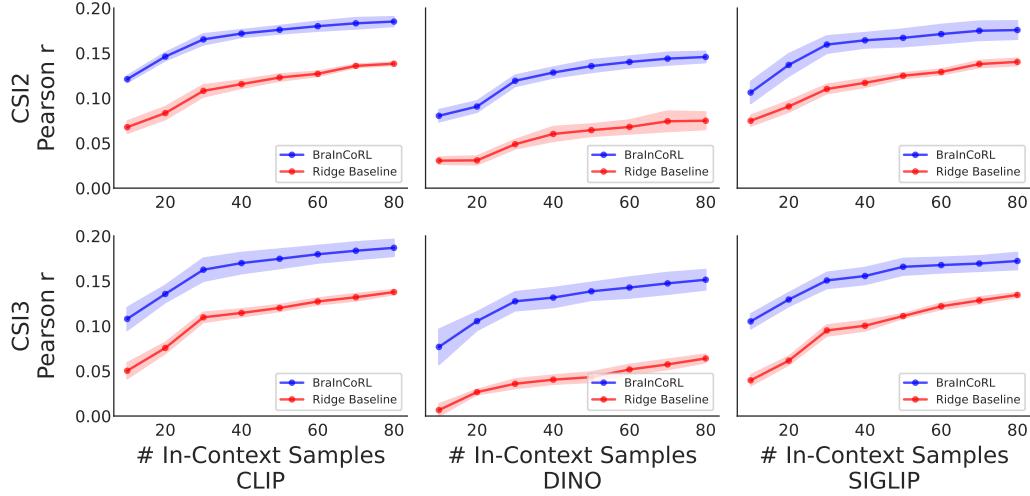
Figure S.13: **Support-set size versus voxelwise Pearson r in BOLD5000.** Each panel shows the mean voxelwise Pearson correlation between predicted and actual BOLD5000 responses for BraInCoRL and ridge regression, plotted against the number of in-context samples.

A.11 Dimensional reduction of predicted response function weights on more subjects

In this section, we utilize UMAP to visualize the BraInCoRL-predicted voxelwise weights under the CLIP image encoding backbone for subject S1-S8. The cortical maps show color-coded mappings that align well with functionally-defined regions.
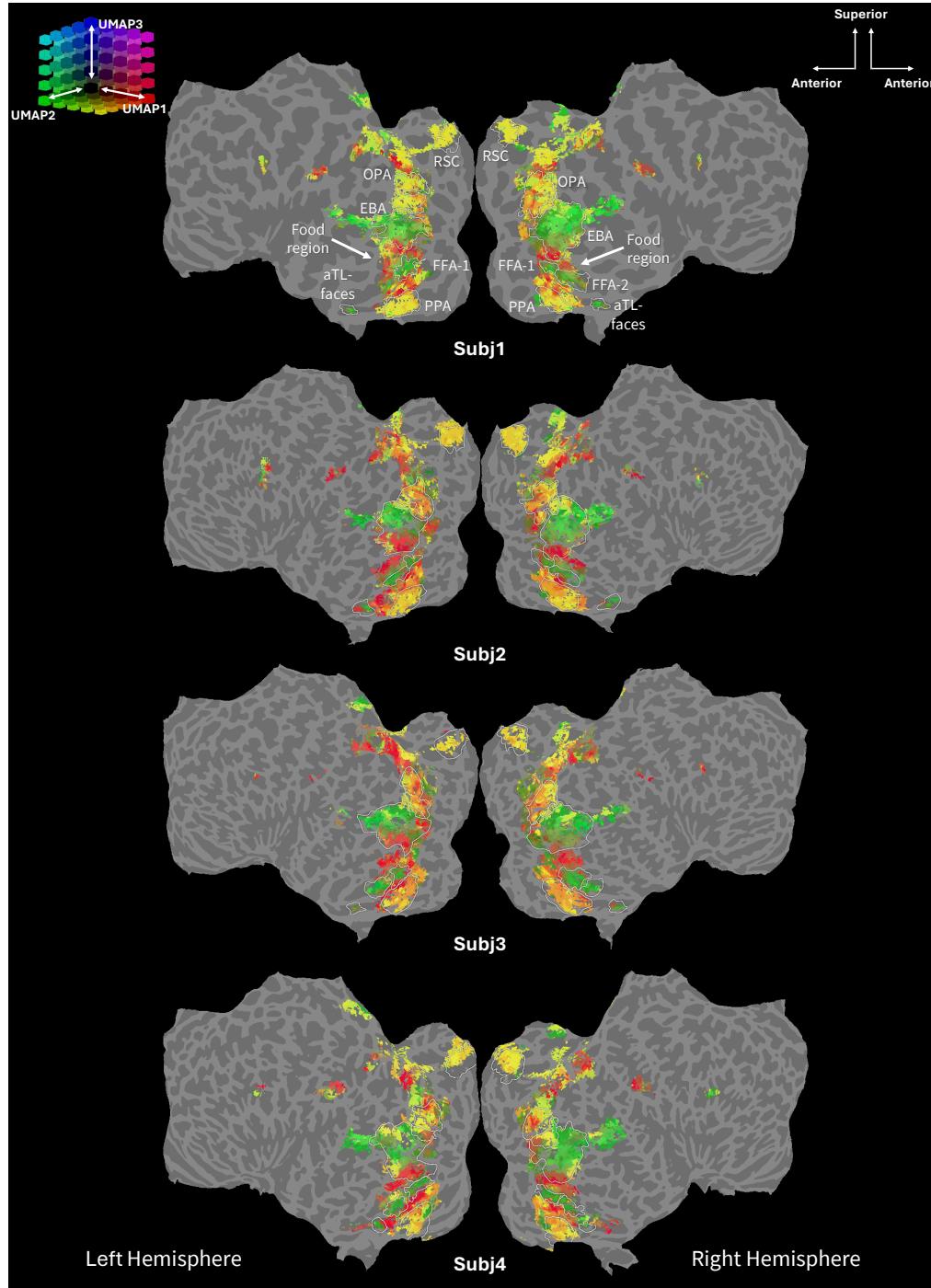

Figure S.14: **Dimensional reduction of predicted response weights for subject S1-S4 under CLIP backbone.** The cortical maps show color-coded mappings that align well with functionally-defined regions: body and face regions (EBA and FFA/aTL-faces), place regions (RSC/OPA/PPA), and food regions (in red).
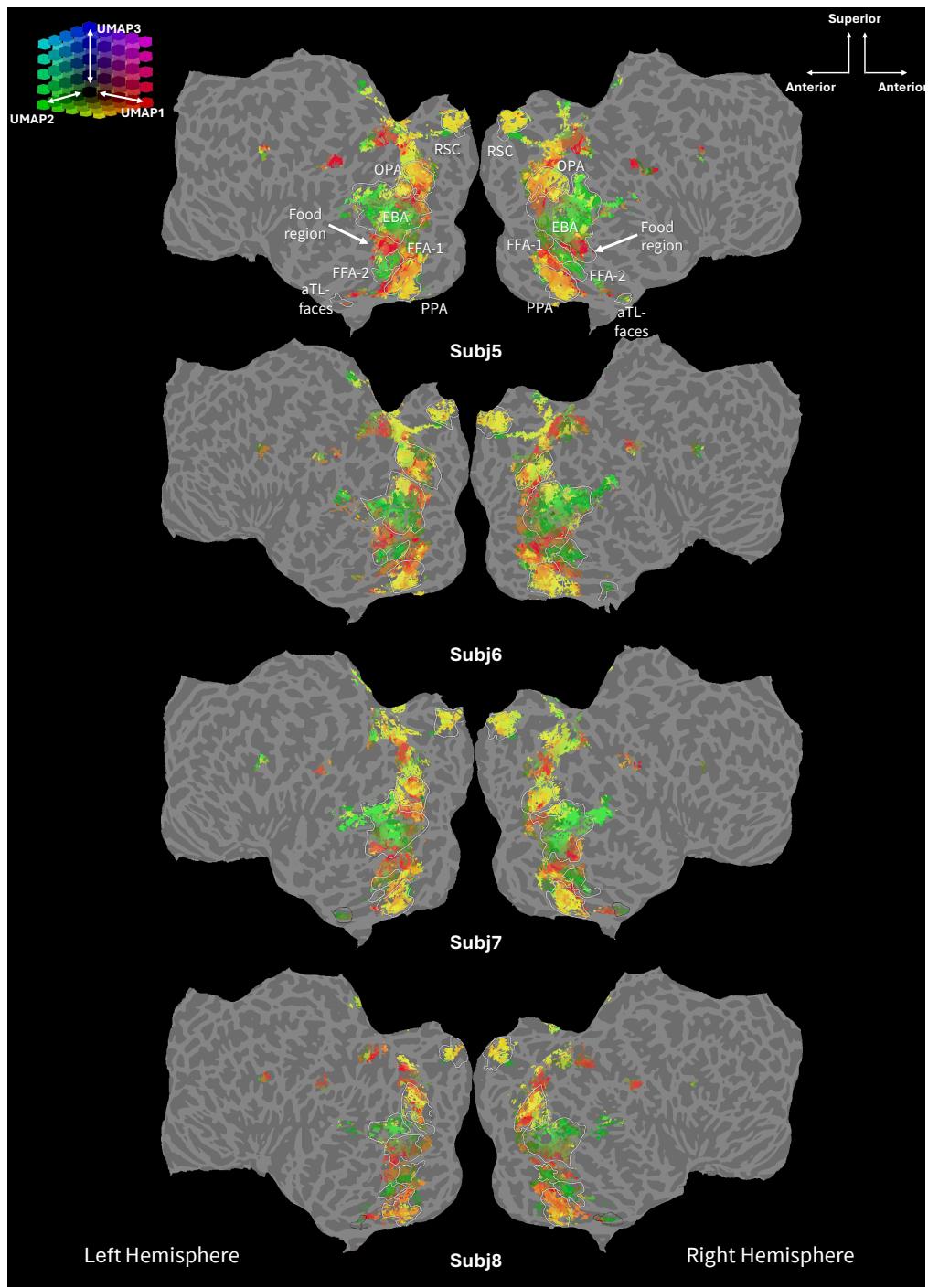
Figure S.15: **Dimensional reduction of predicted response weights for subject S5-S8 under CLIP backbone.** The cortical maps show color-coded mappings that align well with functionally-defined regions: body and face regions (EBA and FFA/aTL-faces), place regions (RSC/OPA/PPA), and food regions (in red).

A.12 Predicting cortical responses from natural language prompts on more subjects

In this section, we further predict cortical responses from natural language prompts on subject 2-8. For each semantic category, we convert a natural language prompt into a CLIP text embedding, project it into the image feature space, and use BraInCoRL to predict the corresponding voxel activation map. The predicted activations align closely with known t-statistic of category-selective region, illustrating the potential for zero-shot, language-driven functional mapping of visual cortex.
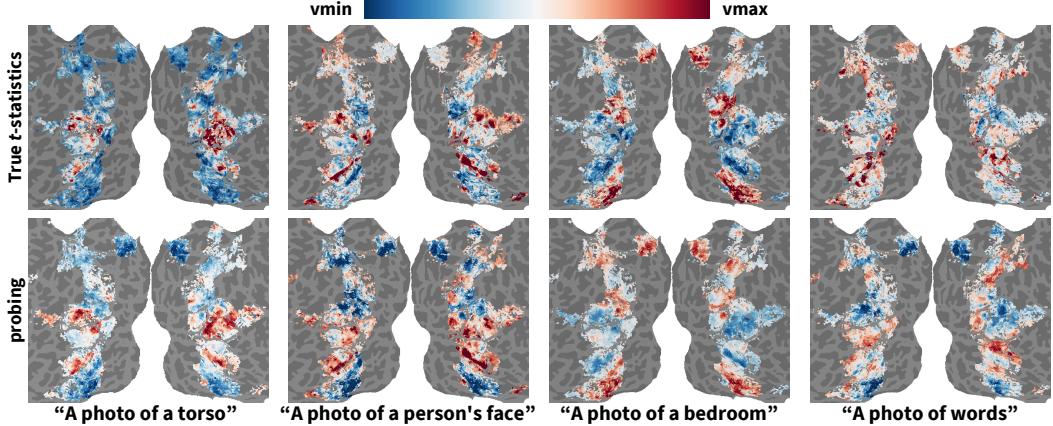
Figure S.16: **Predicting responses of natural language prompts for subject 2.** We convert each text prompt corresponding to a semantic category into a CLIP text embedding, project it into image-feature space, and predict its cortical activation on subject 2. The resulting activation maps closely match the t-statistics of known category-selective regions, demonstrating the feasibility of language-driven, zero-shot functional mapping of the visual cortex.
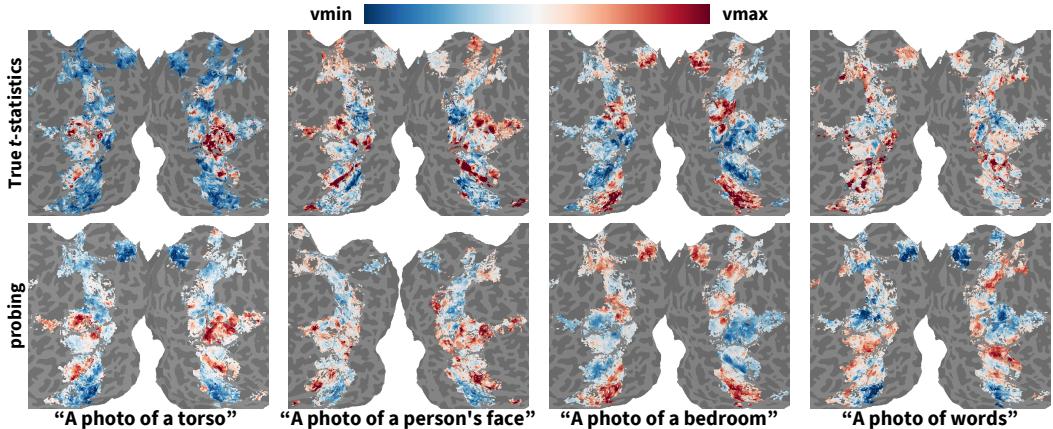

Figure S.17: **Predicting responses of natural language prompts for subject 3.** We convert each text prompt corresponding to a semantic category into a CLIP text embedding, project it into image-feature space, and predict its cortical activation on subject 3. The resulting activation maps closely match the t-statistics of known category-selective regions, demonstrating the feasibility of language-driven, zero-shot functional mapping of the visual cortex.
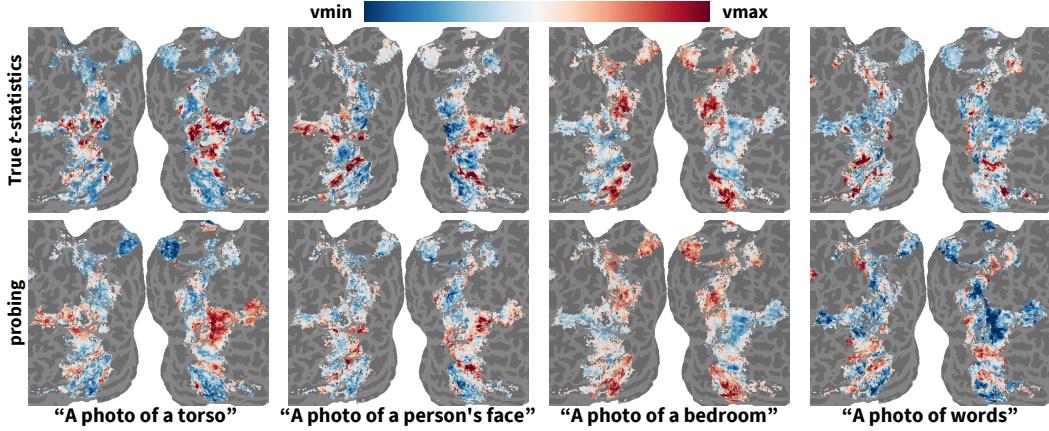
Figure S.18: **Predicting responses of natural language prompts for subject 4.** We convert each text prompt corresponding to a semantic category into a CLIP text embedding, project it into image-feature space, and predict its cortical activation on subject 4. The resulting activation maps closely match the *t*-statistics of known category-selective regions, demonstrating the feasibility of language-driven, zero-shot functional mapping of the visual cortex.
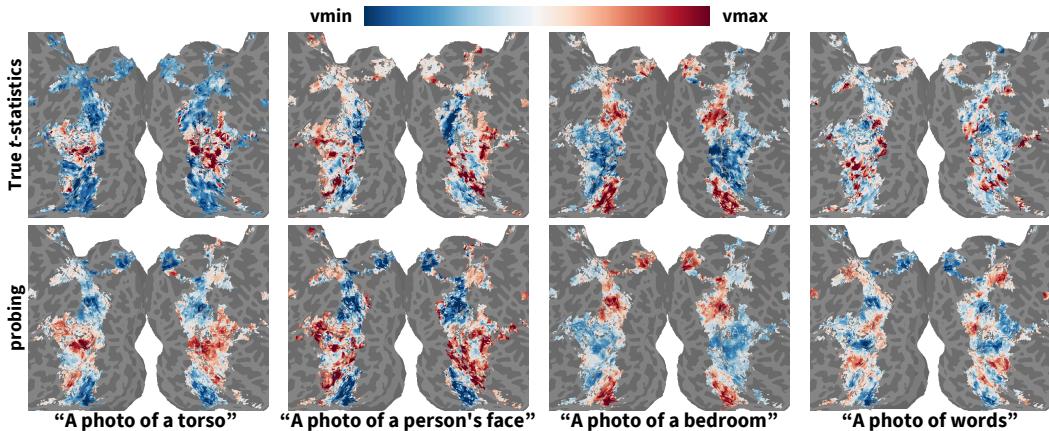

Figure S.19: **Predicting responses of natural language prompts for subject 5.** We convert each text prompt corresponding to a semantic category into a CLIP text embedding, project it into image-feature space, and predict its cortical activation on subject 5. The resulting activation maps closely match the *t*-statistics of known category-selective regions, demonstrating the feasibility of language-driven, zero-shot functional mapping of the visual cortex.
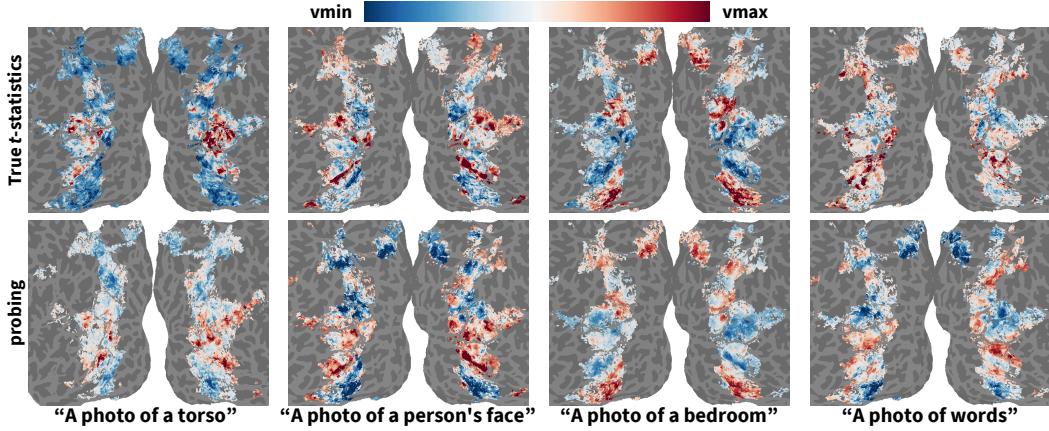
Figure S.20: **Predicting responses of natural language prompts for subject 6.** We convert each text prompt corresponding to a semantic category into a CLIP text embedding, project it into image-feature space, and predict its cortical activation on subject 6. The resulting activation maps closely match the *t*-statistics of known category-selective regions, demonstrating the feasibility of language-driven, zero-shot functional mapping of the visual cortex.
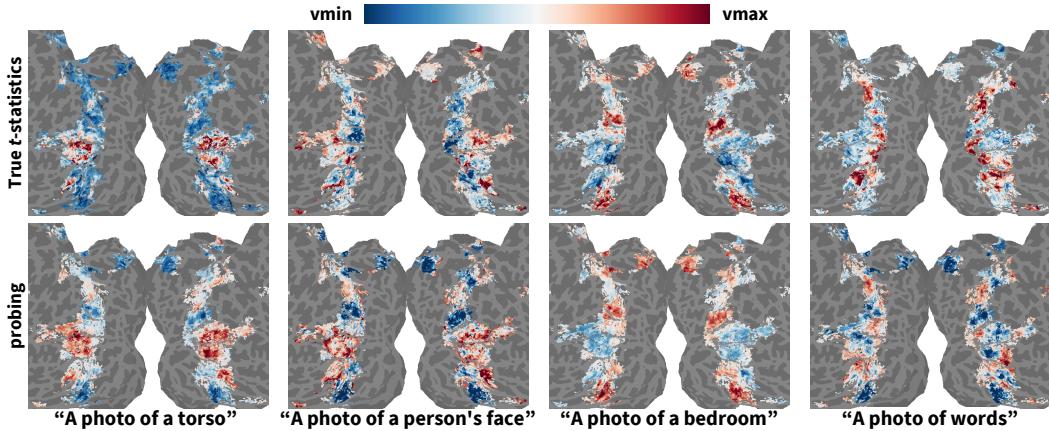

Figure S.21: **Predicting responses of natural language prompts for subject 7.** We convert each text prompt corresponding to a semantic category into a CLIP text embedding, project it into image-feature space, and predict its cortical activation on subject 7. The resulting activation maps closely match the *t*-statistics of known category-selective regions, demonstrating the feasibility of language-driven, zero-shot functional mapping of the visual cortex.
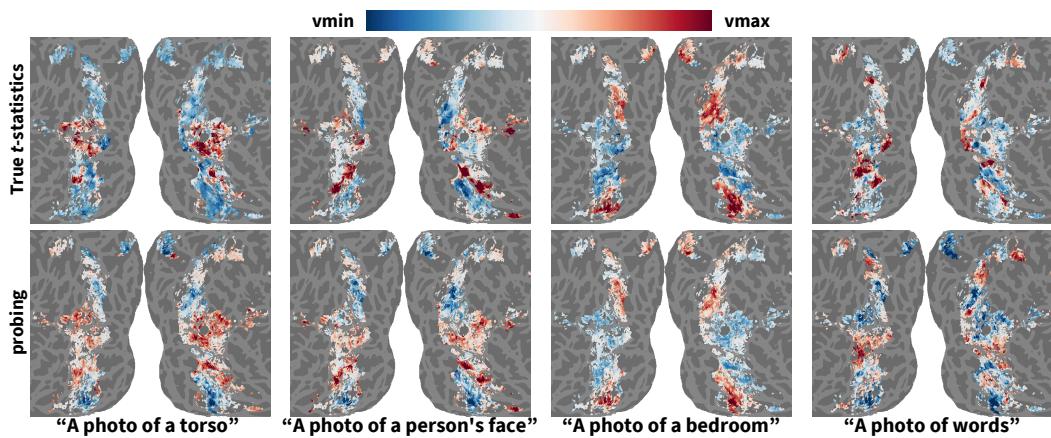
Figure S.22: **Predicting responses of natural language prompts for subject 8.** We convert each text prompt corresponding to a semantic category into a CLIP text embedding, project it into image-feature space, and predict its cortical activation on subject 8. The resulting activation maps closely match the *t*-statistics of known category-selective regions, demonstrating the feasibility of language-driven, zero-shot functional mapping of the visual cortex.

A.13 Voxelwise prompt classification accuracy for more subjects

In this section, we further quantify the semantic specificity of BraInCoRL's voxelwise predictions on subject 3-8. we compute, for each subject and each category-selective ROI, the fraction of voxels whose peak predicted activation corresponded to the semantic category named by the text prompt.

Table S.13: **Voxelwise prompt classification accuracy for subjects 3 and 4.** Each cell shows the percentage of voxels in a given category selective region (columns) whose peak predicted activation was elicited by a specific semantic prompt (rows, see Appendix) for subject 3 and 4. Using only 100 support images, BraInCoRL effectively localizes category-selective regions with high data efficiency.

	Bodies		Faces		Places		Food		Words	
	S3	**S4**	**S3**	**S4**	**S3**	**S4**	**S3**	**S4**	**S3**	**S4**
Bodies	**0.57**	**0.42**	0.23	0.12	0.04	0.03	0.16	0.20	0.12	0.19
Faces	0.29	0.36	**0.60**	**0.66**	0.02	0.03	0.11	0.05	0.20	0.12
Places	0.04	0.08	0.02	0.06	**0.84**	**0.82**	0.15	0.20	0.08	0.16
Food	0.07	0.09	0.14	0.12	0.09	0.09	**0.53**	**0.51**	**0.51**	**0.43**
Words	0.02	0.04	0.01	0.04	0.01	0.02	0.05	0.05	0.09	0.10

Table S.14: **Voxelwise prompt classification accuracy for subjects 5 and 6.** Each cell shows the percentage of voxels in a given category selective region (columns) whose peak predicted activation was elicited by a specific semantic prompt (rows, see Appendix) for subject 5 and 6. Using only 100 support images, BraInCoRL effectively localizes category-selective regions with high data efficiency.

	Bodies		Faces		Places		Food		Words	
	S5	**S6**	**S5**	**S6**	**S5**	**S6**	**S5**	**S6**	**S5**	**S6**
Bodies	**0.54**	**0.64**	0.17	0.21	0.01	0.08	0.19	0.15	0.27	0.25
Faces	0.29	0.25	**0.65**	**0.63**	0.00	0.04	0.03	0.05	0.20	0.15
Places	0.06	0.01	0.05	0.01	**0.88**	**0.65**	0.13	0.09	0.10	0.04
Food	0.09	0.07	0.13	0.11	0.10	0.20	**0.64**	**0.66**	**0.39**	**0.41**
Words	0.02	0.03	0.01	0.04	0.00	0.04	0.01	0.06	0.05	0.15

Table S.15: **Voxelwise prompt classification accuracy for subjects 7 and 8.** Each cell shows the percentage of voxels in a given category selective region (columns) whose peak predicted activation was elicited by a specific semantic prompt (rows, see Appendix) for subject 7 and 8. Using only 100 support images, BraInCoRL effectively localizes category-selective regions with high data efficiency.

	Bodies		Faces		Places		Food		Words	
	S7	**S8**	**S7**	**S8**	**S7**	**S8**	**S7**	**S8**	**S7**	**S8**
Bodies	**0.69**	**0.57**	0.26	0.15	0.01	0.07	0.08	0.20	0.19	0.17
Faces	0.19	0.25	**0.59**	**0.59**	0.01	0.04	0.04	0.10	0.22	0.18
Places	0.04	0.05	0.02	0.03	**0.89**	**0.58**	0.12	0.06	0.13	0.05
Food	0.06	0.10	0.11	0.20	0.07	0.26	**0.68**	**0.57**	**0.32**	**0.47**
Words	0.02	0.03	0.03	0.03	0.01	0.04	0.07	0.08	0.14	0.12

A.14 Additional evaluation of BraInCoRL on NSD dataset

In this section, we provide two more evaluation metrics, namely Pearson R, and Spearman's rank correlation coefficient (Spearman's ρ) for NSD dataset on Subject 1. In this case, the BraInCoRL model has not been trained or finetuned on Subject 1, while the Fully Trained model is trained on 9,000 images from this subject.
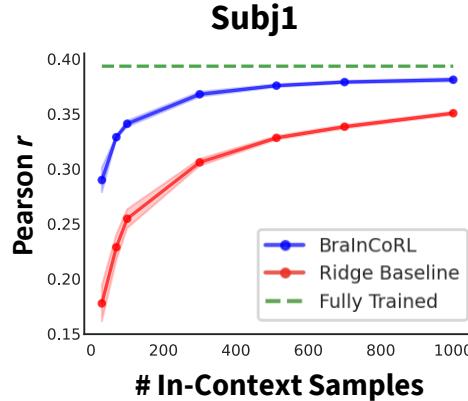

Figure S.23: Voxel-wise Pearson R for BraInCoRL, within-subject ridge regression baseline and fully-trained reference model (NSD dataset, CLIP backbone, Subject 1, higher is better).
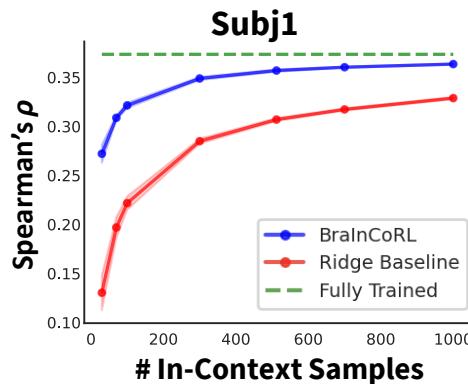
Figure S.24: Voxel-wise Spearman's ρ for BraInCoRL, within-subject ridge regression baseline and fully-trained reference model (NSD dataset, CLIP backbone, Subject 1, higher is better).

A.15 Additional evaluation of BraInCoRL on BOLD5000 dataset

In this section, we provide two more evaluation metrics, namely explained variance and Spearman's rank correlation coefficient (Spearman's ρ) for BOLD5000 dataset on Subject CSI1.
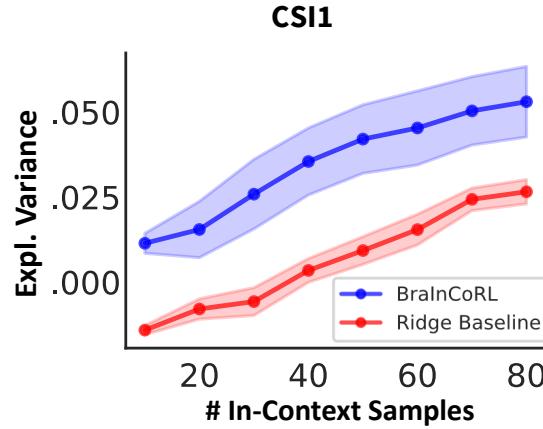

Figure S.25: Voxel-wise explained variance for BraInCoRLand within-subject ridge regression baseline (BOLD5000 dataset, CLIP backbone, Subject CSI1, higher is better).
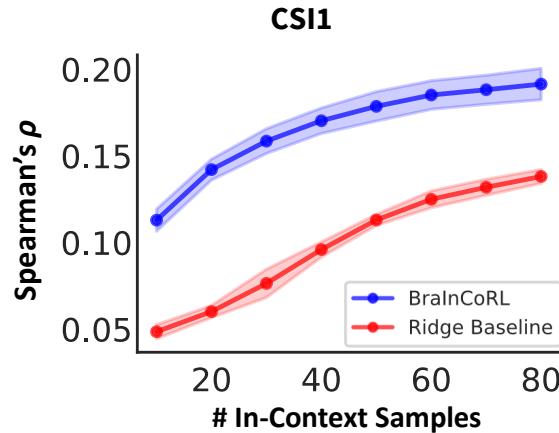

Figure S.26: Voxel-wise Spearman's ρ for BraInCoRLand within-subject ridge regression baseline (BOLD5000 dataset, CLIP backbone, Subject CSI1, higher is better).

A.16 Evaluation of each training stage's contribution

In this section, we present the voxelwise explained variance for BraInCoRLtrained until different training stages. The results show that the progression from synthetic foundation → context flexibility → biological adaptation ensures that each fundamental challenge, namely response function coverage, variable context handling, and biological realism, is systematically addressed in the optimal order.
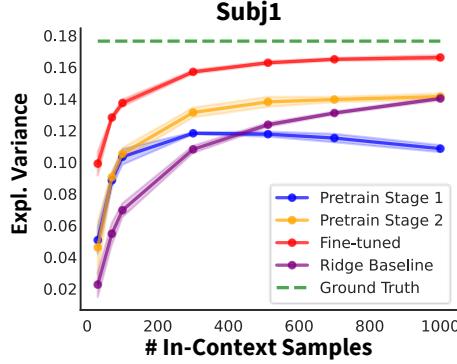

Figure S.27: Voxelwise explained variance for BraInCoRLwith CLIP backbone for NSD Subject 1 on different training stages, compared with ridge baseline and fully trained reference model (higher is better).
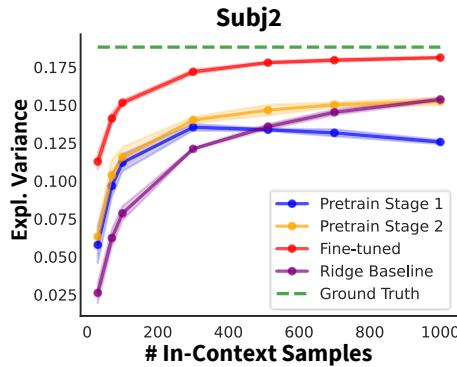

Figure S.28: Voxelwise explained variance for BraInCoRLwith CLIP backbone for NSD Subject 2 on different training stages, compared with ridge baseline and fully trained reference model (higher is better).

A.17 Performance of BraInCoRL conditioned on the full 9000-image set

In this section, we evaluate the performance of BraInCoRL with 9,000 in-context samples, for the four NSD subjects (S1, S2, S5, S7) focues by our main paper. The performance difference is less than 1% across all subjects when compared to a fully trained model (which is fit to converge on each subject's entire 9,000-image training set using gradient descent over multiple epochs). This means BraInCoRL achieves 94-99% of the fully trained model's performance.

Table S.16: Voxel-wise explained variance of BraInCoRL with the CLIP backbone compared with the fully trained reference model. The difference variance explained is less than 1%.

Method	Subject			
	S1	S2	S5	S7
Fully Trained	0.1765	0.1882	0.2310	0.1554
BrainCoRL	**0.1667**	**0.1817**	**0.2225**	**0.1541**

A.18 Evaluation on the choice of loss function during training

In this section, we conducted an ablation study on the choice of different loss functions during the finetuning stage optimization and evaluated the model performance.

Our experimental results show that MSE and L1 losses achieve similar performance across all context sizes, with minimal differences. This suggests that both metrics are equally effective for capturing voxelwise neural response patterns.

In addition, the hybrid loss of $0.5 \times \text{MSE loss} + 0.5 \times (1 - \text{cosine similarity})$ underperforms by approximately 2-4% compared to MSE/L1. We argue this is because although cosine similarity captures directional relationships between predicted and true responses, this additional constraint limit the model's ability to accurately predict response magnitudes.
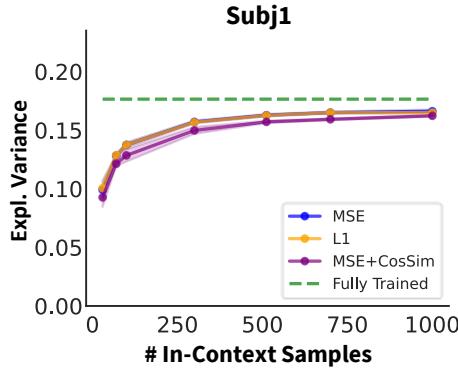

Figure S.29: Voxel-wise explained variance of different training losses with the CLIP backbone for Subject 1 (higher is better).

A.19 Ablation on logit scaling

In this section, we provide an additional ablation study to evaluate the effect of the logit-scaling, where we report the Voxel-wise explained variance of our BrainCoRL model and the model with the exact same structure but without logit scaling. It is shown that the logit scaling significantly boosts the model performance and generizability of various in-context support set sizes.
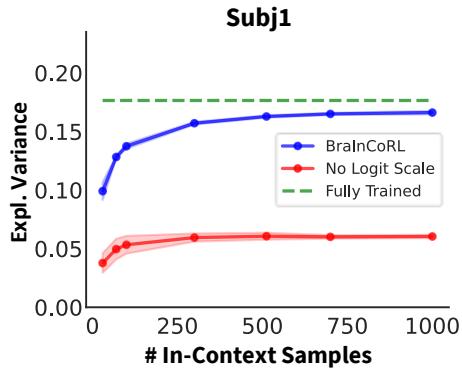

Figure S.30: Voxel-wise explained variance for BrainCoRL with CLIP backbone for Subj 1, compared to the same model architecture but without logit scaling (higher is better).

Appendix References

[1] Jianlin Su. Analyzing the scale operation of attention from the perspective of entropy invariance, Dec 2021. URL https://kexue.fm/archives/8823.

[2] David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. *arXiv preprint arXiv:2202.12172*, 2022.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. 4(7):2025, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/,checkedon.