BrainAlign: Leveraging EEG Foundation Models for Symmetric, Interpretable Alignment with Visual Representations

Vijay Jayawant Harkare
Department of Computer Science
CIMS, New York University
New York, NY 10012
vijay.harkare@nyu.edu

Lakshminarayan Subramanian
Department of Computer Science
CIMS, New York University
New York, NY 10012
lakshmi@cs.nyu.edu

Abstract

Understanding how the human brain represents visual objects is a fundamental challenge that can be addressed by aligning brain activity recordings in the form of electroencephalography (EEG) recordings with features from computer vision models. However, prior work has predominantly relied on custom EEG encoders trained on limited, task-specific data, which restricts their ability to learn generalizable, brain-like representations. In this work, we propose an alternative approach, moving from task-specific encoders to a representation-first approach. We leverage a large-scale pretrained EEG foundation model, CBraMod, to provide a rich and robust foundation for learning brain-aligned representations. We introduce BrainAlign, a contrastive learning framework that uses a brain-inspired projection network to align EEG representations with those from various image encoders (ResNet50, CORNet-S, and CLIP). To evaluate the quality of these aligned representations, we test our framework on the challenging 200-way zero-shot visual object classification task. Using a CORNet-S image encoder, BrainAlign achieves a top-1 accuracy of 14.2%, exceeding the NICE framework's baseline and performing comparably to state-of-the-art methods that use only vision and EEG modalities. Furthermore, our framework demonstrates significant computational efficiency, reducing the required training epochs by 70% compared to training from scratch. Moreover, analysis of the learned representational geometry reveals a structure consistent with established phenomena of the human visual system. Collectively, these results in performance, computational efficiency, and biological plausibility validate our representation-first approach, highlighting the potential of foundation models to bridge the gap between neural and artificial representations.

1 Introduction

Aligning neural activity with representations from computational models is a fundamental approach to understanding the principles of brain function. This endeavor not only advances our basic scientific knowledge but also holds immense potential for transformative applications, particularly in developing next-generation Brain-Computer Interfaces (BCIs) for clinical and consumer use[24, 15]. Among non-invasive neuroimaging methods, electroencephalography (EEG) is a promising modality due to several key advantages. Its high temporal resolution captures neural dynamics at the millisecond scale, aligning with the rapid nature of visual processing, while its portability and low cost make it ideal for practical, real-world applications outside of laboratory settings[24, 25]. In contrast, modalities such as fMRI, despite offering superior spatial resolution, are limited by poor temporal dynamics and expensive, cumbersome hardware[22].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Data on the Brain & Mind.

Historically, the utility of EEG for decoding was hindered by low signal-to-noise ratios and flawed paradigms like block-design experiments, which introduced temporal confounds[24, 27]. The field has since shifted toward more robust methodologies, with the Rapid Serial Visual Presentation (RSVP) paradigm and large-scale datasets like THINGS-EEG2 enabling the study of neural responses to thousands of natural images[6]. This evolution led to self-supervised contrastive learning emerging as the dominant approach for aligning the high-dimensional space of EEG signals with rich visual representations[24]. However, a critical limitation pervades these modern methods: they almost exclusively rely on custom EEG encoders trained from scratch on a single alignment task. This methodology is fundamentally constrained, as an encoder optimized solely for one task is unlikely to learn the generalizable, brain-like neural codes that capture the full richness of brain activity. To overcome this, we propose an alternative "representation-first" approach that leverages the power of EEG foundation models[2]. These models, pre-trained on massive and diverse neural datasets, learn universal and robust representations that serve as a superior starting point. By fine-tuning from this rich representational base, we can learn alignments that are more data-efficient, performant, and, crucially, more likely to be biologically plausible[8, 26].

To rigorously evaluate the quality of the learned representations, we utilize the 200-way zero-shot visual object classification task. This task serves as a challenging benchmark for two reasons: First, its zero-shot nature directly tests the model's ability to generalize to unseen semantic concepts, a key indicator of a robustly learned representation space. Second, it is an established evaluation paradigm within the BCI and neuro-AI communities[5, 23, 24], allowing for direct comparison with prior state-of-the-art methods. Success on this task, therefore, is not an end in itself, but a strong proxy for the quality and generalizability of the underlying brain-visual alignment.

To implement this representation-first approach, we introduce BrainAlign, a framework designed for the symmetric and interpretable alignment of EEG and visual representations. While leveraging a foundation model addresses the primary challenge of learning robust neural codes, our framework is also designed to investigate several other critical gaps in existing research. First, unlike architecturally asymmetric models, BrainAlign is designed to be bi-directional, capturing the reciprocal nature of information processing in the brain[28, 18]. Second, we move beyond "black box" models by incorporating methods that enhance mechanistic interpretability, allowing us to use the model as a scientific instrument. Finally, we address the open question of which visual feature space best aligns with EEG signals. By systematically comparing a purely hierarchical model (ResNet[7]), a brain-inspired recurrent model (CORNet-S[13]), and a vision-language model (CLIP[16]), we can probe the nature of the optimal visual-neural alignment.

This paper introduces a framework for visual object classification from EEG that directly addresses the aforementioned gaps. Our contributions can be summarized as follows:

- We introduce BrainAlign, a framework that operationalizes a representation-first approach by leveraging a state-of-the-art EEG foundation model (CBraMod[26]) to learn robust and generalizable neural representations for alignment with visual features.
- We systematically investigate the nature of visual-neural alignment by contrastively aligning these powerful EEG representations with three distinct and neuroscientifically motivated visual backbones: a purely visual hierarchical model (ResNet50)[7], a brain-inspired recurrent model (CORNet-S)[13], and a vision-language model (CLIP)[19]. This comparative analysis allows for an examination of the resulting representational geometry.
- We demonstrate the bi-directional symmetry of the shared representation space learned via contrastive alignment. This provides the basis for future work on encoding stimuli into brain-like representations, thus allowing the investigation of various neuro-scientific hypotheses.
- We assess the framework's interpretability by visualizing learned importance weights corresponding to distinct brain regions within the ventral visual pathway.
- We analyze the quality of the shared representation space through its intrinsic information content and its performance on downstream tasks.

2 Related work

Aligning neural and computational models. The effort to map visual representations in the brain has progressed from early fMRI studies, which established that object categories could be decoded from cortical activity[24], to modern electrophysiological methods like EEG. The high temporal resolution of EEG is better suited to capture the rapid dynamics of visual perception[2]. A significant methodological advance was the adoption of the Rapid Serial Visual Presentation (RSVP) paradigm, which, combined with large-scale datasets, enabled the field to move beyond simple classification to ambitious zero-shot decoding tasks using deep learning[6, 9]. This research now largely falls under the broader goal of integrative benchmarking, where computational models are quantitatively evaluated on their ability to predict neural and behavioral data, a practice formalized by platforms like Brain-Score[21].

Contrastive learning for EEG-vision alignment. The current state-of-the-art for aligning EEG signals with visual features is self-supervised contrastive learning[14]. The pioneering NICE framework demonstrated that a contrastive loss could effectively map EEG and image embeddings (e.g., from CLIP) into a shared space for zero-shot recognition[23]. While language-guided extensions like NICE++ have shown performance gains by using textual descriptions to refine the alignment[24], they do so by introducing a third modality (language). As our work is focused on the fundamental principles of direct EEG-vision alignment, we compare against uni-modal visual encoders. Subsequent work has introduced sophisticated refinements to address challenges such as the "modality gap". For instance, BraVL uses a multimodal VAE to learn a unified latent space[5], VE-SDN introduces a semantic decoupling module to align only the shared information[3], and others leverage guidance from large language models to refine the alignment[24]. A common thread, however, unites these advanced methods: they all train their EEG encoders from scratch for a specific alignment task. This approach is fundamentally limited, as the encoders must simultaneously learn basic neural feature extraction and high-level semantic alignment, a challenge that our work directly addresses.

EEG foundation models. These models are pre-trained on massive and diverse EEG corpora, such as the TUH-EEG dataset[17], to learn universal, robust, and generalizable representations of brain activity. Architectures like BENDR[10] and LaBraM[8] established the viability of this approach. We employ CBraMod[26], a state-of-the-art foundation model whose criss-cross transformer architecture is uniquely suited to capturing the spatio-temporal dynamics of EEG. By starting with these rich, pre-trained representations, we reframe the problem from one of end-to-end training to one of targeted fine-tuning. This aligns with a broader movement in computational neuroscience away from purely predictive "black box" models and toward models that are mechanistically interpretable[11]. The goal is to build transparent, falsifiable models of neural computation, where the internal workings can be causally linked to behavior and brain activity. Our representation-first approach, grounded in a powerful foundation model, is a critical step in this direction.

3 Method

The methodology of this study is designed to validate our central thesis: that leveraging a pre-trained EEG foundation model provides a more robust and biologically plausible pathway to learning brain-aligned representations than training task-specific encoders from scratch. To this end, we introduce BrainAlign, a framework designed for the symmetric and interpretable alignment of EEG and visual features. Our experimental design adheres to a subject-dependent paradigm. This choice is rooted in the principle of biological plausibility; as each human brain possesses unique functional characteristics, developing subject-specific models is essential for capturing genuine neural representations, rather than learning a non-representative 'average' brain model. In this section, we will detail the architecture of the BrainAlign framework (refer Figure 1), the rationale behind its components, and the contrastive learning procedure used for training.

3.1 BrainAlign architecture

The BrainAlign framework consists of two parallel processing streams—an EEG branch and an image branch—that learn to project their respective outputs into a shared representation space. The EEG branch is designed to address the fundamental limitations of conventional approaches that

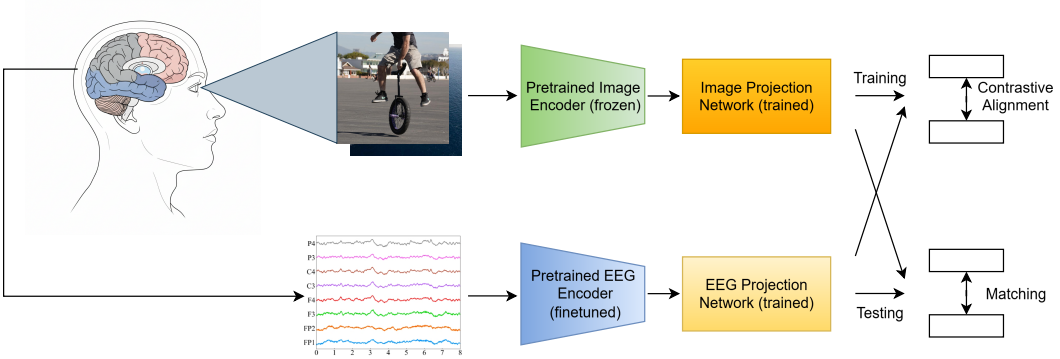

Figure 1: The BrainAlign framework for EEG foundation model-based object classification. The framework relies on powerful pretrained EEG and image encoders, and while finetuning the EEG encoder, trains the projection networks using contrastive learning to align the representation spaces from both branches. Testing is done by matching EEG branch representations with pre-obtained image branch templates for test images.

train encoders from scratch. Such methods are not only computationally expensive (e.g., up to 200 epochs[23]) but also risk learning brittle, task-specific representations, as they must learn low-level features and high-level alignment simultaneously. Our framework circumvents this by utilizing a pre-trained EEG foundation model, CBraMod[26], as the encoder. By starting with the rich, general-purpose representations learned from diverse datasets[8, 10, 17], our model can achieve high performance with substantially less fine-tuning. Following this encoder, we introduce a custom projection network designed with strong neuroscientific priors of regional cortical processing. The architecture adopts a multi-stream design that segregates channels into functionally distinct groups (occipital, parietal, temporal, and global) and integrates them via a learnable gating mechanism, yielding a functionally grounded and interpretable embedding. The detailed mathematical formulation of this regional aggregation process is provided in Appendix A.

A central scientific question of this study is what kind of computational visual feature space aligns most effectively with neural representations. To investigate this, the image branch of our framework is designed to be modular. We systematically compare three distinct, neuroscientifically motivated image encoders, each representing a different hypothesis about visual processing: a hierarchical feedforward model (ResNet50), a brain-inspired recurrent model (CORNet-S), and a multimodal vision-language model (CLIP). This comparative experiment is therefore designed not simply to find the best-performing model, but to use alignment performance as evidence to adjudicate between these competing computational theories of visual representation. A detailed description of each of these encoders is available in Appendix A. Following the selected encoder, a simple 2-layer MLP with GeLU activation serves as a projection network to map the image features into the shared representation space.

3.2 Contrastive learning

The core of the training process is to align EEG and image features in a shared embedding space. This is achieved using a symmetric contrastive loss function, similar to the one introduced in CLIP. The symmetric nature of this loss is critical, as it encourages the learned latent space to be bi-directionally informative. This ensures that an EEG representation can be used to identify its corresponding image (decoding) and, equally, that an image representation can identify its EEG counterpart (encoding), a property essential for building models that reflect the brain's reciprocal processing pathways (refer Appendix A).

Given a mini-batch of N paired EEG and image samples, we first extract their respective feature vectors, f_e and f_i, using the EEG and image encoders. These features are then projected into a shared embedding space of dimension D by projection heads P_{eeg} and P_{img}.

The projected features for the k-th sample are denoted as $z_e^{(k)} = P_{eeg}(f_e^{(k)})$ and $z_i^{(k)} = P_{img}(f_i^{(k)})$. These features are L2-normalized:

$$\hat{z}_e^{(k)} = \frac{z_e^{(k)}}{\left\| z_e^{(k)} \right\|_2} \quad \text{and} \quad \hat{z}_i^{(k)} = \frac{z_i^{(k)}}{\left\| z_i^{(k)} \right\|_2}$$

The similarity between the j-th EEG feature vector and the k-th image feature vector in the batch is calculated as the cosine similarity (dot product of normalized vectors), scaled by a learnable temperature parameter τ:

$$s_{jk} = \tau \cdot \left\langle \hat{z}_e^{(j)}, \hat{z}_i^{(k)} \right\rangle$$

The objective is to maximize the similarity of corresponding pairs (where $j = k$) while minimizing it for all other non-corresponding pairs within the batch. This is framed as a classification problem using the cross-entropy loss (refer Appendix A). The loss is calculated symmetrically for both EEG-to-image and image-to-EEG directions.

The loss for predicting the correct image pairing for a given EEG signal is:

$$\mathcal{L}_{\text{eeg}} = -\frac{1}{N} \sum_{j=1}^{N} \log \frac{\exp(s_{jj})}{\sum_{k=1}^{N} \exp(s_{jk})}$$

Similarly, the loss for predicting the correct EEG pairing for a given image is:

$$\mathcal{L}_{\text{img}} = -\frac{1}{N} \sum_{j=1}^{N} \log \frac{\exp(s_{jj})}{\sum_{k=1}^{N} \exp(s_{kj})}$$

The final training objective is the average of these two losses:

$$\mathcal{L}_{\text{total}} = \frac{\mathcal{L}_{\text{eeg}} + \mathcal{L}_{\text{img}}}{2}$$

4 Experimental setup and results

4.1 Dataset, preprocessing, and quality analysis

4.1.1 Dataset and preprocessing

For this study, we selected the THINGS-EEG2[6] dataset due to its neuroscientific validity and high temporal resolution. This dataset contains EEG responses from 10 subjects viewing natural images presented using a rapid serial visual presentation (RSVP) paradigm. The RSVP protocol is designed to elicit stimulus-specific neural responses while minimizing contributions from higher-order cognitive processes, making the data suitable for training models on object recognition. The dataset comprises 82,160 trials across 16,740 unique image conditions, which map to 1,854 object classes. We adhere to the original study's split, using 1,654 classes for training and 200 classes for the zero-shot evaluation task. For the test set, one image per class was selected for the 200-way classification task. EEG data was recorded from 64 channels using an EASYCAP system, out of which 63 were recording channels and one was stimulus channel.

We followed standard EEG preprocessing steps, consistent with those applied by Song et al.. The raw data was epoched into 1000 ms trials post-stimulus onset and baseline-corrected using the mean of the 200 ms pre-stimulus period. A bandpass filter was applied to retain frequencies between 0.1 and 100 Hz. For all analyses, the data was down-sampled from 1000 Hz to 250 Hz, and multivariate noise normalization was performed to reduce correlated noise across channels. This frequency was chosen in accordance with the Nyquist-Shannon sampling theorem. All trial repetitions for each image condition were averaged to increase the signal-to-noise ratio. During training, the EEG data was further down-sampled to 200 Hz to match the input requirements of the CBraMod foundation model. For the image branch, we utilized pre-computed image representations from ResNet50, CORNet-S,

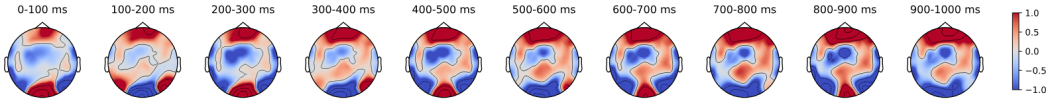

Figure 2: Topographical maps of EEG responses from one subject averaged over all training image conditions across 10 time intervals.

Table 1: A comparison of different model performances (top-1 accuracies) across 10 subjects for the EEG-to-Image 200-way zero-shot classification task

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Ave	Std
BraVL[5]	6.1	4.9	5.6	5.0	4.0	6.0	6.5	8.8	4.3	7.0	5.8	1.4
NICE[23]	12.3	10.4	13.1	16.4	8.0	14.1	15.2	20.0	13.3	14.9	13.8	3.3
NICE-GA[23]	15.2	13.9	14.7	17.6	9.0	16.4	14.9	20.3	14.1	19.6	15.6	3.2
CBraMod (finetuned) + CLIP	14.5	9.5	14.0	11.5	10.0	19.0	11.5	16.5	13.5	17.0	13.7	3.1
CBraMod (finetuned) + ResNet-50	12.0	12.0	12.0	9.5	9.0	21.5	12.0	16.0	10.0	18.5	13.2	4.1
CBraMod (finetuned) + CORNet-S	11.5	13.0	13.5	16.0	10.0	20.5	14.5	14.0	12.5	16.5	14.2	2.9
CBraMod (frozen) + CLIP	2.5	5.0	7.0	7.5	2.5	6.5	5.0	7.0	4.5	10.0	5.7	2.3
CBraMod (frozen) + ResNet-50	5.0	5.5	6.5	4.5	6.0	9.0	5.0	10.0	2.5	6.5	6.0	2.2
CBraMod (frozen) + CORNet-S	4.0	6.5	7.0	5.5	6.0	8.5	5.5	7.5	2.5	9.0	6.2	2.0

and CLIP, as provided by the original dataset creators and Song et al., to facilitate faster model training and evaluation.

4.1.2 Quality analysis

We included all 63 channels to ensure our model captures the distributed activity of the ventral stream, extending beyond just the occipital-parietal regions. As shown in Figure 2, the spatiotemporal dynamics confirm a feedforward flow from V1 to anterior temporal areas, justifying the full-montage approach. Please refer to Appendix B for a detailed discussion of these activation patterns.

4.2 Evaluation framework and results

Our experimental investigation centered on two key questions, evaluated on a subject-dependent basis to account for inter-subject variability[20]. First, to test our central hypothesis, we compared two training strategies for the CBraMod encoder: fine-tuning the pre-trained weights versus keeping them frozen. Second, to investigate the nature of the optimal visual feature space, we paired each EEG strategy with the three visual backbones (ResNet50, CORNet-S, and CLIP). This resulted in six model configurations per subject, which were evaluated on the bi-directional 200-way zero-shot classification task (chance-level accuracy: 0.5%). For a deeper, qualitative assessment of the learned representations, we also designed a series of targeted representational analyses (e.g., RSA, time-resolved encoding). A detailed description of each of these representational analysis methods is provided in Appendix D.

The performance of our six model configurations was evaluated and compared against the NICE, NICE-GA, and BraVL frameworks[23, 5]. In this work, we focus our primary analysis on top-1 accuracy, as it serves as the most stringent metric for evaluating the quality and "brain-alikeness" of the learned representations. Unlike top-5 accuracy, which allows for a wider margin of error, top-1 accuracy directly probes the model's ability to select the single correct item from 200 distinct choices. This provides a direct measure of the representation's discriminative power—its ability to distinguish between fine-grained concepts from neural data, which is a key characteristic of the brain's own highly specific and efficient visual processing system. The mean top-1 accuracies across all subjects are presented in Table 1 and Table 2; for completeness, top-5 accuracies are provided in Appendix F.

Our primary finding validates the central hypothesis of this work: leveraging a pre-trained foundation model as an inductive bias via fine-tuning is superior to using it as a static feature extractor. As shown in Tables 1 and 2, all fine-tuned models dramatically outperformed their frozen-backbone counterparts. This large and statistically significant improvement in top-1 accuracy ($p < 0.01$, Wilcoxon Signed-Rank test) demonstrates that the fine-tuning process is critical for adapting the

Table 2: A comparison of different model performances (top-1 accuracies) across 10 subjects for the Image-to-EEG 200-way zero-shot classification task

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Ave	Std
CBraMod (finetuned) + CLIP	23.0	17.0	16.0	20.0	17.5	23.0	19.0	26.5	18.5	30.5	21.1	4.6
CBraMod (finetuned) + ResNet-50	17.0	26.5	19.5	22.5	21.0	29.0	15.5	24.5	13.5	29.0	21.8	5.5
CBraMod (finetuned) + CORNet-S	17.0	25.5	21.5	25.0	18.0	33.5	23.0	27.0	16.0	26.0	23.2	5.3
CBraMod (frozen) + CLIP	4.5	7.5	9.5	11.5	8.5	10.5	5.5	13.5	2.5	11.0	8.4	3.4
CBraMod (frozen) + ResNet-50	6.0	10.5	6.5	12.0	10.5	12.0	5.5	13.0	5.5	8.0	8.9	3.0
CBraMod (frozen) + CORNet-S	3.5	10.0	9.5	7.0	9.0	12.0	4.5	12.5	4.0	13.0	8.5	3.6

foundation model's general-purpose features into a highly discriminative semantic space, one that is better suited for the specific task of visual object recognition from EEG. This result strongly supports our representation-first approach.

Having established the importance of fine-tuning, we next investigated which visual feature space aligns best with the adapted EEG representations. Among the fine-tuned models, the configuration using the brain-inspired recurrent CORNet-S encoder achieved the highest average top-1 accuracy in both EEG-to-Image (14.2%) and Image-to-EEG (23.2%) directions. This suggests that its representations, shaped by recurrent connections designed to mimic the primate ventral stream, provide a more suitable target space for alignment with neural data. While not statistically significant ($p > 0.05$), the consistent top performance of CORNet-S provides compelling evidence in favor of using brain-inspired architectures for such alignment tasks.

Our best-performing model (CBraMod fine-tuned + CORNet-S) is highly competitive with current state-of-the-art methods, significantly outperforming BraVL (5.8%) and the base NICE (13.8%) frameworks, and achieving an accuracy comparable to the more complex NICE-GA model (15.6%). Crucially, this performance is achieved with marked computational efficiency. All fine-tuned models converged within 60 epochs, a 70% reduction in training time (measured in terms of the number of epochs). This efficiency is not a trivial improvement; it is a critical factor for the scalability and practical viability of our subject-dependent paradigm. As a new model must be trained for each new subject, a significant reduction in training time directly translates to lower computational costs and a greater capacity to apply the framework to larger participant cohorts.

4.3 Model interpretability and representational plausibility

To assess model interpretability, we visualized the regional importance weights learned by the EEG projection network as a topographical map (Figure 3). The visualization shows that the model consistently assigned higher weights to occipital, parieto-occipital, and inferior temporal channels compared to frontal channels. This learned weight distribution is consistent with the known functional anatomy of the ventral visual pathway, providing evidence for the biological plausibility of the model. Furthermore, the fine-tuned models learned a weight distribution that more closely resembled this neuroscientific prior compared to the frozen-backbone models. This observation provides a potential mechanistic explanation for the performance gap reported in Section 4.2: the fine-tuning process not only adapts the feature space but also enables the model to learn a neuroanatomically correct attention policy, focusing on the most informative brain regions for the task. The superior performance of the fine-tuned models is therefore not just a numerical result, but a consequence of learning a more biologically plausible processing strategy.

To provide deeper evidence for the quality of the learned representations beyond classification accuracy, we conducted a series of representational analyses (see Appendix E for full details). These analyses confirmed three key points. First, time-resolved encoding showed that our aligned representations captured significant, dynamically evolving neural information, mirroring the known temporal progression of the visual (Figure 4). Second, Representational Similarity Analysis (RSA) revealed that the geometry of the space learned by the fine-tuned models had a significantly higher correlation with the brain's own representational geometry (EEG embeddings) compared to the frozen models (Appendix Figure 5). Third, high accuracy on cross-modal retrieval tasks confirmed that the space is robustly bi-directional. Taken together, these results provide converging evidence that the performance gains from our foundation model framework are rooted in its ability to learn a shared latent space that is more structurally and dynamically aligned with the brain's internal representations.

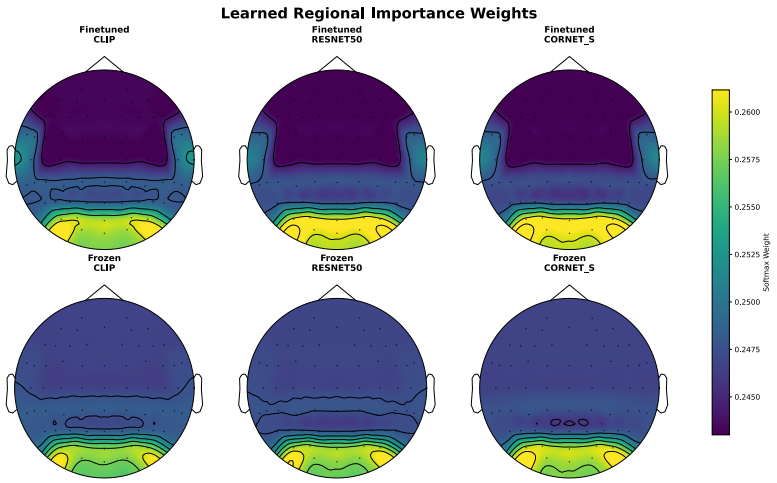

Figure 3: Topographical map of brain region importance weights learned by the EEG projection network.

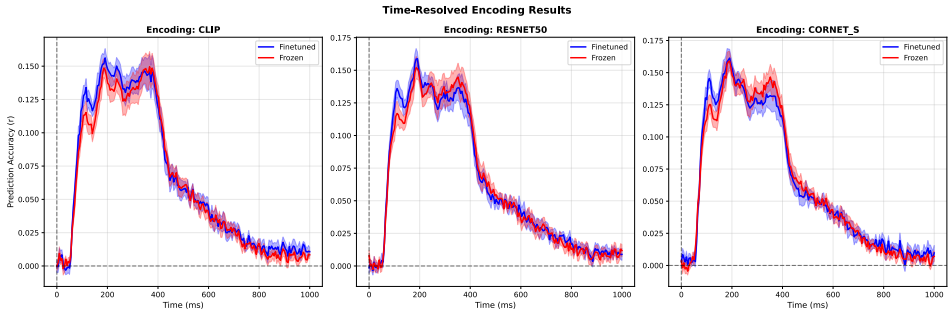

Figure 4: Prediction accuracy of raw EEG signals from image representations using time-resolved encoding models.

5 Conclusion and future work

In this work, we have demonstrated that leveraging pre-trained EEG foundation models via fine-tuning constitutes a more powerful, efficient, and biologically plausible paradigm for aligning neural and artificial visual representations. Our BrainAlign framework achieves competitive performance on the challenging 200-way zero-shot classification benchmark while drastically reducing the required training time by 70%. Crucially, this strong quantitative performance is underpinned by qualitative evidence of greater neuroscientific validity: interpretability analyses reveal that our fine-tuned model learns a neuroanatomically correct attentional policy, while representational similarity analyses confirm that its learned geometry is more congruent with the brain's own. These findings collectively establish the "representation-first" approach as a robust and scientifically informative path forward, paving the way for the development of more sophisticated BCIs and more transparent computational models of brain function.

Limitations. All results are based on subject-dependent models, and therefore, cross-subject generalization remains to be explored yet. The 200-way zero-shot classification task,while a good and commonly-used proxy for measuring quality of alignment, leaves actual downstream task performance on tasks like image reconstruction to future work. While we tried to establish interpretability in various ways, large-scale user studies are required to demonstrate the biological plausibility of the model, which is beyond the scope of this study.

Acknowledgments and Disclosure of Funding

The authors thank Dr. Cristina Savin for providing detailed feedback. We also acknowledge New York University's Department of Computer Science for providing access to essential resources, including MATLAB. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] P. Bao, L. She, M. McGill, and D. Y. Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020.

[2] M. Berto. Eeg foundation models: Unlocking the next generation of neurotechnology. `https://www.brainaccess.ai/eeg-foundation-models-unlocking-the-next-generation-of-neurotechnology/`. Accessed: 02 September 2025.

[3] H. Chen, L. He, Y. Liu, and L. Yang. Visual neural decoding via improved visual-eeg semantic consistency. *arXiv preprint arXiv:2408.06788*, 2024.

[4] J. J. DiCarlo and D. D. Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

[5] C. Du, K. Fu, J. Li, and H. He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

[6] A. T. Gifford, K. Dwivedi, G. Roig, and R. M. Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119754. URL `https://www.sciencedirect.com/science/article/pii/S1053811922008758`.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] W.-B. Jiang, L.-M. Zhao, and B.-L. Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

[9] Z. Jiao, H. You, F. Yang, X. Li, H. Zhang, and D. Shen. Decoding eeg by visual-guided deep neural networks. In *IJCAI*, volume 28, pages 1387–1393. Macao, 2019.

[10] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

[11] J. W. Krakauer, A. A. Ghazanfar, A. Gomez-Marin, M. A. MacIver, and D. Poeppel. Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93(3):480–490, Feb 2017. ISSN 0896-6273. doi: 10.1016/j.neuron.2016.12.041. URL `https://doi.org/10.1016/j.neuron.2016.12.041`.

[12] N. Kriegeskorte, M. Mur, and P. A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

[13] J. Kubilius, M. Schrimpf, K. Kar, R. Rajalingham, H. Hong, N. Majaj, E. Issa, P. Bashivan, J. Prescott-Roy, K. Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.

[14] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1): 857–876, 2021.

[15] X.-Y. Liu, W.-L. Wang, M. Liu, M.-Y. Chen, T. Pereira, D. Y. Doda, Y.-F. Ke, S.-Y. Wang, D. Wen, X.-G. Tong, et al. Recent applications of eeg-based brain-computer-interface in the medical field. *Military Medical Research*, 12(1):14, 2025.

[16] Z. Lu and Y. Wang. Category-selective neurons in deep networks: Comparing purely visual and visual-language models. *arXiv preprint arXiv:2502.16456*, 2025.

[17] I. Obeid and J. Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

[18] K. Qiao, J. Chen, L. Wang, C. Zhang, L. Zeng, L. Tong, and B. Yan. Category decoding of visual stimuli from human brain activity using a bidirectional recurrent neural network to simulate bidirectional information flows in human visual cortices. *Frontiers in neuroscience*, 13:692, 2019.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[20] S. Saha and M. Baumert. Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: a review. *Frontiers in computational neuroscience*, 13:87, 2020.

[21] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, K. Schmidt, D. L. K. Yamins, and J. J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2020. doi: 10.1101/407007. URL `https://www.biorxiv.org/content/early/2020/01/02/407007`.

[22] D. Sharon, M. S. Hämäläinen, R. B. Tootell, E. Halgren, and J. W. Belliveau. The advantage of combining meg and eeg: comparison to fmri in focally stimulated visual cortex. *Neuroimage*, 36(4):1225–1235, 2007.

[23] Y. Song, B. Liu, X. Li, N. Shi, Y. Wang, and X. Gao. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*, 2023.

[24] Y. Song, Y. Wang, H. He, and X. Gao. Recognizing natural images from eeg with language-guided contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[25] A. Trafton. In the blink of an eye. `https://news.mit.edu/2014/in-the-blink-of-an-eye-0116`. Accessed: 01 September 2025.

[26] J. Wang, S. Zhao, Z. Luo, Y. Zhou, H. Jiang, S. Li, T. Li, and G. Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024.

[27] L. Xu, M. Xu, T.-P. Jung, and D. Ming. Review of brain encoding and decoding mechanisms for eeg-based brain–computer interface. *Cognitive neurodynamics*, 15(4):569–584, 2021.

[28] Y. Zhang, Y. Wang, M. Azabou, A. Andre, Z. Wang, H. Lyu, T. I. B. Laboratory, E. Dyer, L. Paninski, and C. Hurwitz. Neural encoding and decoding at scale. *arXiv preprint arXiv:2504.08201*, 2025.

A Architectural and model details

A.1 EEG projection network formulation

The process for deriving the aggregated EEG vector from the output of the EEG encoder, $F = \{f_1, f_2, \ldots, f_C\}$, is as follows. The channels are grouped into four disjoint sets based on their location: occipital (C_O), parietal (C_P), temporal (C_T), and other (C_{Other}). For each region $R \in \{O, P, T, Other\}$, the features are first averaged:

$$\bar{f}_R = \frac{1}{|C_R|} \sum_{c \in C_R} f_c$$

This mean-pooled feature vector is then passed through a region-specific projection network P_R:

$$f'_R = P_R(\bar{f}_R)$$

The model learns a set of importance weights, $\mathbf{w} = [w_O, w_P, w_T, w_{Other}]$, which are derived from a learnable parameter parameter vector $\mathbf{v}$ via the softmax function:

$$\mathbf{w} = \text{softmax}(\mathbf{v})$$

Finally, the weighted features from each region are concatenated to form the final aggregated EEG feature vector, z_{agg}:

$$z_{\text{agg}} = [w_O \cdot f'_O \oplus w_P \cdot f'_P \oplus w_T \cdot f'_T \oplus w_{Other} \cdot f'_{Other}]$$

where $\oplus$ denotes the concatenation operation.

A.2 Image encoder details

We systematically compare three distinct image encoders, each representing a different hypothesis about visual processing.

ResNet50[7] This model represents the 'hierarchical feedforward' hypothesis, where visual information is processed through a series of increasingly complex, feedforward layers. Its alignment performance serves as a baseline for a standard, highly-performant computer vision architecture.

CORNet-S[13] This model represents the 'brain-inspired recurrence' hypothesis. It was explicitly designed to model the primate ventral visual stream and incorporates recurrent connections, which are a key feature of the visual cortex. Its performance tests whether an architecturally more brain-like model yields better alignment.

CLIP[19] This model represents the 'semantic embedding' hypothesis. Pre-trained on image-text pairs, its representations are not purely visual but are deeply structured by language and semantics. Its performance probes whether the brain's representation of objects is more akin to a rich, multimodal semantic space than a purely visual one.

A.3 Additional details on contrastive learning

The loss function has been deliberately chosen to be a symmetric contrastive loss, to reflect the brain's reciprocal processing pathways. This acknowledges the symmetric processes of imagining images by decoding some latent representations and interpreting the visual information obtained through the optic nerve by encoding it in some latent space, which continuously occur in the brain.

Furthermore, we also clarify the framing of the self-supervised learning problem as a classification problem. It is important to note that throughout the training process, no class labels have been utilized. The classification is done based on similarity between EEG and vision encoder representations for the concerned mini-batch. In the context of this architecture, classification implies finding an EEG

Table 3: Hyperparameter settings used for model training.

Name	Value
Batch size	1024
Learning rate	0.0002
Adam β_1	0.5
Adam β_2	0.999
Logit scale (τ)	$\log(1/0.07)$
Projection dimension (EEG and Image)	800
EEG encoder embedding dimension	800
Image encoder embedding dimension (CLIP)	784
Image encoder embedding dimension (CORNet-S and ResNet50)	3000
Dropout (all layers)	0.2
Validation split size	740 samples
Training split size	16540 samples
Test split size	200 samples

representation for the image, and symmetrically, finding an image for a given EEG representation. Since all processing occurs strictly with the available data, without utilizing any labels, the objective remains a valid self-supervised learning objective.

B Data quality and channel selection

While prior work has sometimes restricted analysis to 17 occipital and parietal channels, we retained all 63 channels for model training, similar to Song et al[23]. This decision is motivated by the fact that the ventral visual pathway, which is critical for object recognition, extends beyond the occipital and parietal lobes into the inferior temporal cortex[1]. Including all channels allows the model to potentially capture a more complete representation of the distributed neural activity underlying visual processing.

To confirm the data quality across these channels, we performed a temporal and spatial analysis of the EEG responses. As shown in the main text (Figure 2), the activation patterns are consistent with established neuroscientific findings: an initial increase in activity in the occipital lobe (0-100 ms), followed by propagation to the temporal lobe. This characterizes feedforward processing along the ventral visual stream, including V1, V2, V3, PIT, CIT and AIT areas, thus validating the suitability of the full 63-channel dataset.

C Hyperparameter choices

The hyperparameters used for training all models are provided in Table 3[23].

D Representational analysis methods

To gain deeper insight into the structure and biological plausibility of the shared latent space, we conducted a series of targeted representational analyses, as described below.

Quality of neural information content To verify that the aligned image representations captured meaningful neural information, we performed a time-resolved encoding analysis. Using a nested cross-validated Ridge regression model, we predicted EEG signals at each time point from the static image features of the aligned space. High prediction accuracy in this analysis would indicate that the contrastive learning process successfully embedded neurally-relevant visual features into the representations, validating the Image-to-EEG mapping.

Similarity to brain's representational geometry To assess the biological plausibility of the learned space, we compared its internal structure to that of the brain using time-resolved Representational Similarity Analysis (RSA)[12]. We computed Representational Dissimilarity Matrices (RDMs) for

the model and for the neural data at each time point. A high correlation between the model and brain RDMs over time would demonstrate that our framework learns a representational geometry that dynamically mirrors the brain's own processing trajectory.

Bi-directional symmetry and alignment Finally, to evaluate the overall alignment and bi-directional utility of the final shared space, we conducted two analyses. First, a static RSA measured the global alignment between the final EEG and image representational geometries. Second, a cross-modal retrieval task directly tested the framework's symmetry by evaluating its ability to retrieve the correct EEG vector from its image counterpart, and vice-versa. Success in these tasks is a direct measure of how well the two modalities were fused into a coherent, symmetric representational space.

E Results of representational analyses

Figure 5 shows the results of various representational analyses.

E.1 Analysis of temporal dynamics in raw EEG data

The first set of analyses evaluated the extent to which the learned image representations in the shared space captured the temporal dynamics of the raw neural signals. Figure 4 (time-resolved encoding) and 5A (time-resolved RSA) show that the ability to predict or correlate with the raw EEG signal peaks between 100-250 ms and remains significant until around 600 ms post-stimulus. This temporal profile is highly consistent with the known hierarchical progression of feedforward processing along the human ventral visual stream[4].

Notably, the performance between the fine-tuned and frozen model paradigms is largely comparable in these analyses. This finding is significant: it suggests that the large-scale pre-training of the CBraMod foundation model is sufficient to learn and preserve the core, low-level temporal dynamics of visual neural processing. This validates the use of the foundation model as a strong starting point, as it provides a robust neuro-temporal prior before any task-specific adaptation occurs.

E.2 Analysis of the aligned shared representation space

The second set of analyses assessed a different, more central question: the quality of the final, shared representational space created by the contrastive learning process. Instead of comparing to raw EEG, these analyses directly measure the geometric alignment between the final EEG representations and the image representations.

The results, shown in Figures 5B, 5C, and 5D, provide unequivocal evidence for our central hypothesis. The representational alignment, as measured by RSA correlation, is dramatically and statistically significantly higher in the fine-tuned paradigm compared to the frozen paradigm (Figure 5C, $p<0.001$). This demonstrates that while the frozen backbone provides a strong temporal prior, it is insufficient for creating a high-fidelity shared semantic space. The process of fine-tuning is critical; it allows the model to adapt the general-purpose neural features into representations that are specifically and geometrically aligned with their visual counterparts. The high absolute correlation values and cross-modal retrieval accuracies (Figure 5B) for the fine-tuned models further confirm the overall effectiveness of the BrainAlign framework in learning a robust, bi-directionally useful shared space.

F Additional results

Tables 4 and 5 present additional top-5 accuracy results for the EEG-to-Image and Image-to-EEG 200-way zero shot classification tasks respectively.

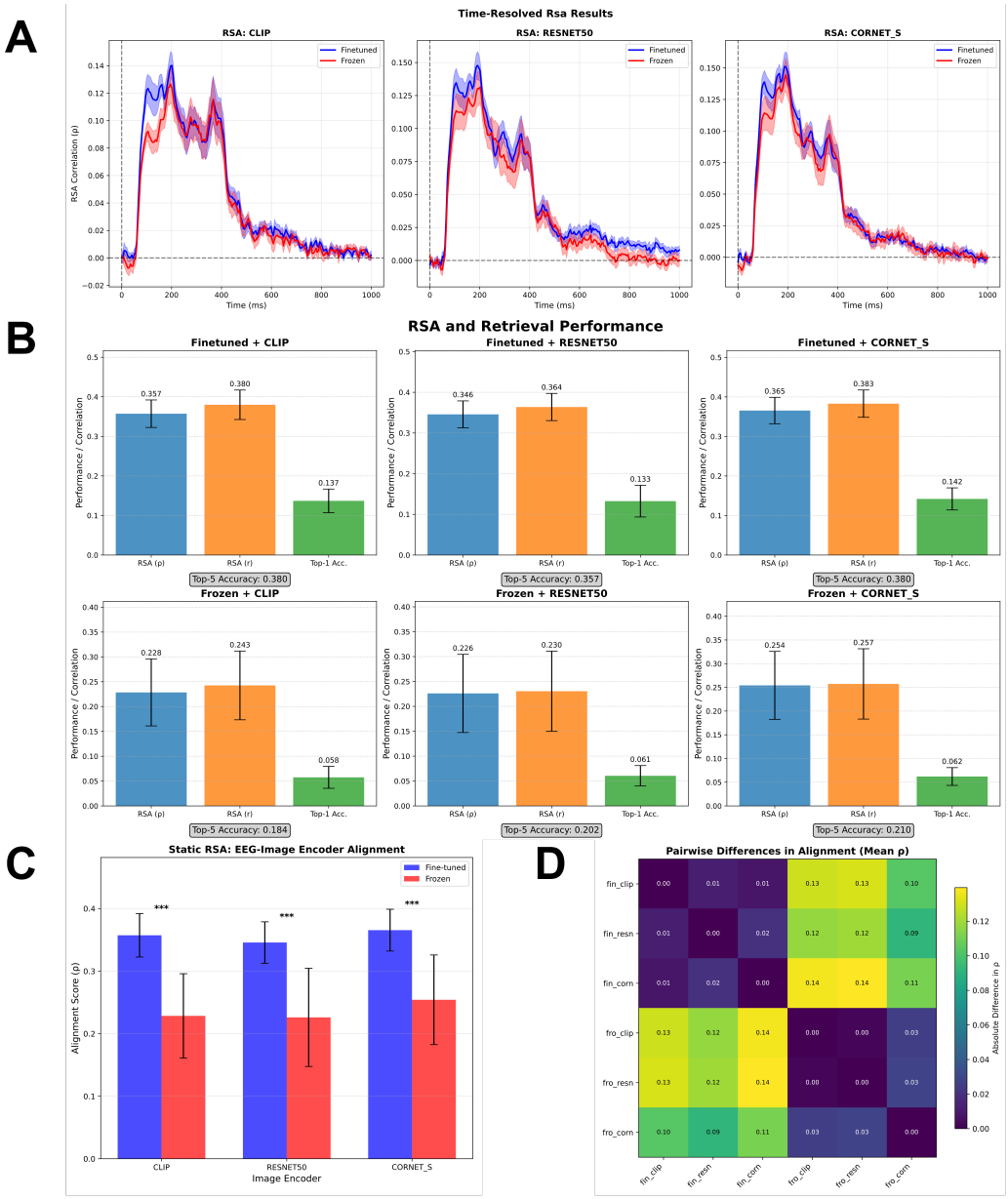

Figure 5: Results of representational analyses. (A) RSA correlation of raw EEG signals with image representations using time-resolved RSA analysis. (B) Mean Pearson (ρ) and Spearman (r) coefficients for RSA between EEG and image representations for all subjects, along with top-1 and top-5 EEG-to-Image retrieval accuracies across model configurations. (C) Comparison of EEG-Image representation alignment between fine-tuned and frozen paradigms using RSA between EEG and Image representations averaged over all subjects (*** indicates statistical significance of $p < 0.001$). (D) Heatmap of pairwise differences in RSA alignment across all model configurations.

Table 4: A comparison of different model performances (top-5 accuracies) across 10 subjects for the EEG-to-Image 200-way zero-shot classification task

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Ave	Std
BraVL[5]	17.9	14.9	17.4	15.1	13.4	18.2	20.4	23.7	14.0	19.7	17.5	3.2
NICE[23]	36.6	33.9	39.0	47.0	26.9	40.6	42.1	49.9	37.1	41.9	39.5	6.5
NICE-GA[23]	40.1	40.1	42.7	48.9	29.7	44.4	43.1	52.1	39.7	46.7	42.8	6.1
CBraMod (finetuned) + CLIP	37.0	30.5	37.0	31.0	29.5	49.5	36.0	44.0	39.0	46.5	38.0	6.9
CBraMod (finetuned) + ResNet-50	29.0	34.0	34.0	29.0	30.5	52.0	29.0	47.0	31.5	41.0	35.7	8.2
CBraMod (finetuned) + CORNet-S	31.0	39.0	36.0	40.5	24.5	50.5	37.5	41.5	32.0	47.0	37.9	7.7
CBraMod (frozen) + CLIP	12.5	16.5	19.0	24.5	13.0	22.5	14.5	22.0	12.5	27.0	18.4	5.4
CBraMod (frozen) + ResNet-50	18.0	17.0	18.5	20.0	18.5	29.5	17.0	26.0	15.0	23.0	20.2	4.5
CBraMod (frozen) + CORNet-S	17.0	22.0	24.5	25.0	21.0	25.0	18.5	23.0	12.0	22.0	21.0	4.1

Table 5: A comparison of different model performances (top-5 accuracies) across 10 subjects for the Image-to-EEG 200-way zero-shot classification task

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Ave	Std
CBraMod (finetuned) + CLIP	54.0	45.5	50.5	55.5	45.0	58.0	51.0	58.5	51.5	60.5	53.0	5.3
CBraMod (finetuned) + ResNet-50	42.5	54.5	48.5	47.0	47.0	60.5	47.0	58.5	45.0	55.0	50.5	6.1
CBraMod (finetuned) + CORNet-S	49.0	57.5	47.5	53.0	43.0	67.0	52.5	64.5	49.5	63.5	54.7	8.1
CBraMod (frozen) + CLIP	15.0	25.5	25.5	33.5	24.5	28.5	17.5	31.0	18.5	30.5	25.0	6.2
CBraMod (frozen) + ResNet-50	17.5	25.0	21.0	27.5	27.5	37.5	20.5	37.5	18.5	32.0	26.4	7.4
CBraMod (frozen) + CORNet-S	20.0	30.5	27.5	32.5	27.0	28.5	24.5	34.5	15.5	36.0	27.6	6.4