Bridging the Semantic Latent Space between Brain and Machine: Similarity Is All You Need

Jiaxuan Chen[1], **Yu Qi**[2, 3, 1*], **Yueming Wang**[4], **Gang Pan**[3, 1, 2]

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2]MOE Frontier Science Center for Brain Science and Brain-Machine Integration, Zhejiang University, Hangzhou, China
[3]The State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China
[4]Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, China
{jiaxuan_chen, qiyu, ymingwang, gpan}@zju.edu.cn

Abstract

How our brain encodes complex concepts has been a long-standing mystery in neuroscience. The answer to this problem can lead to new understandings about how the brain retrieves information in large-scale data with high efficiency and robustness. Neuroscience studies suggest the brain represents concepts in a locality-sensitive hashing (LSH) strategy, *i.e.*, similar concepts will be represented by similar responses. This finding has inspired the design of similarity-based algorithms, especially in contrastive learning. Here, we hypothesize that the brain and large neural network models, both using similarity-based learning rules, could contain a similar semantic embedding space. To verify that, this paper proposes a functional Magnetic Resonance Imaging (fMRI) semantic learning network named BrainSem, aimed at seeking a joint semantic latent space that bridges the brain and a Contrastive Language-Image Pre-training (CLIP) model. Given that our perception is inherently cross-modal, we introduce a fuzzy (one-to-many) matching loss function to encourage the models to extract high-level semantic components from neural signals. Our results claimed that using only a small set of fMRI recordings for semantic space alignment, we could obtain shared embedding valid for unseen categories out of the training set, which provided potential evidence for the semantic representation similarity between the brain and large neural networks. In a zero-shot classification task, our BrainSem achieves an 11.6% improvement over the state-of-the-art.

Introduction

Human visual system actively reorganizes incoming sensory data to allow that people can recognize thousands of complex objects quickly and easily (Yamins and DiCarlo 2016). Nevertheless, the fine-grained relationships between brain activities and visual stimuli are poorly understood (Popham et al. 2021). How our brain encodes properties of concepts is still a central quest of neuroscience. Over the past few decades, decoding visual information from functional Magnetic Resonance Imaging (fMRI) signals has been investigated in different contexts, *e.g.*, classification (Kamitani and Tong 2005; Horikawa and Kamitani 2017; Du et al. 2023), and reconstruction (Du et al. 2022; Gaziv et al. 2022; Chen et al. 2023a,b; Fang, Zheng, and Pan 2023). However, due to the brain's internal cross-modal nature and the absence of effective biological guidance, extracting semantic components from limited fMRI recordings remains challenging, particularly in terms of generalizing to novel categories.

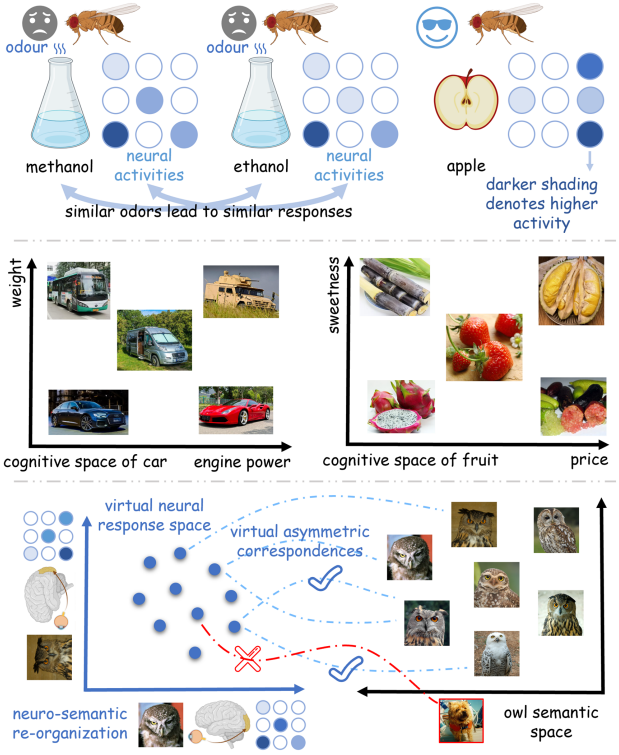

Figure 1: Top: The fruit fly circuit assigns similar neural response patterns to similar odors. Middle: Schematic of two cognitive spaces spanned by object properties. Bottom: The main idea of the proposed BrainSem, which cast the neural representation learning into a fuzzy visual-neural matching paradigm in a self-supervised manner.

Neural code mechanisms suggest that the brain assigns similar neural activity patterns to similar concepts (Bellmund et al. 2018; Stevens 2015). For example, a study (Dasgupta, Stevens, and Navlakha 2017), which investigated the

*Corresponding author.

fly's neural circuit in the odor representation, showed that the fly circuit assigned similar neural codes to similar odors, with an LSH code, as illustrated in Fig. 1 top. For another, Bellmund *et al*. (2018) also have shown evidence that overlapping population responses lead to nearby positions for similar stimuli, and larger distances between dissimilar stimuli in cognitive space, as exhibited in Fig. 1 middle. The way of locality-sensitive encoding can facilitate that behaviors learned from one stimulus can be applied when a similar input is experienced, which can be critical in similarity-driven semantic recognition and reasoning. Recent developments in large pre-training models, *e.g.*, CLIP (Radford et al. 2021), also reveal that concepts with similar semantics are close in the embedding space. What should one expect from using a large neural network to understand our brain?

Computational neuroscience communities have made strides in modeling population responses and neural single-units via neural networks, *e.g.*, the hierarchical properties of sensory cortical processing (Yamins and DiCarlo 2016). However, whether there are locality-sensitive homologies driven by semantic similarity between neural and artificial semantic representation spaces has not been adequately explored. Unraveling this promises to shed new light on the combination of biological and artificial intelligence (Wu, Pan, and Zheng 2013). In this study, our main hypothesis is that the brain and large pre-training CLIP model, both using similarity-based learning rules, have a similar latent semantic representation space. To validate that, we propose a novel BrainSem network, combining recent advances in the representational format of cognitive spaces and contrastive self-supervised learning, to seek a joint semantic latent space across the brain and artificial neural networks.

Technically, in our BrainSem, neural semantic representation learning is boiled down to a fuzzy matching paradigm, which formulates a one (fMRI) to many (diverse images within the same class) loss function. This formulation relaxes the requirement of in-depth visual understanding of stimulus images (*e.g.*, color and texture details), and compels the model to prioritize capturing the high-level semantic information. By doing so, we demonstrated that the BrainSem can be applied to fulfil cross-modal semantic alignment from limited fMRI-image annotations and yield both local-consensus and high-performance neural embedding spaces. These results suggested that similarity-based learning could be the key to generalizable semantic representations between the brain and machine. Overall, our contributions can be summarised as follows:

- We introduce a self-supervised brain semantics learner (BrainSem) with a specially designed neuro-semantic reorganization format and fuzzy matching loss, which allows us to learn a joint semantic latent space bridging the brain and CLIP model from limited fMRI recordings.
- With the help of BrainSem, we obtain shared semantic embedding valid for unseen categories and achieve state-of-the-art zero-shot neural classification, which provides potential evidence for the representation similarity between the brain and large neural network models.
- The embedding space learned from BrainSem not only reveals more visible local semantic consensus properties than CLIP embedding space, but also delineates the abiotic and biotic spatial cognitive areas with a clear linear boundary.

Related Work

Visiual Decoding with Brain Signals. Seeking homogeneous semantic information between brain activities and the corresponding visual stimulus contents has long been a sought objective. Early efforts primarily focus on training a classifier to build the relationship between brain activation pattern and the pre-defined labels using fMRI (Kamitani and Tong 2005; Haxby et al. 2001; Van Gerven et al. 2010; Damarla and Just 2013; Yargholi and Hossein-Zadeh 2016) or electroencephalography (EEG) data (Palazzo et al. 2020; Spampinato et al. 2017; Ahmed et al. 2021). However, the classification-based approach is restricted to the decoding of a certain set of categories, which severely limits its applicability. To mitigate these issues, several identification-based approaches (Kay et al. 2008; Horikawa and Kamitani 2017; Akamatsu et al. 2020) have been investigated. These methods commonly fulfill unknown category decoding via characterizing the relationship between brain activity and visual semantic knowledge, such as visual features extracted from CNNs (Horikawa and Kamitani 2017; Akamatsu et al. 2020) or Gabor wavelet filters (Kay et al. 2008). Though the identification on a large set of categories is possible, the major defect they suffer is that the decoding accuracy critically depends on large-scale labeled stimuli-responses data. Therefore, accurately decoding novel image categories with the limited training paired annotations remains a challenge. Recently, solving neural decoding tasks with advanced deep neural networks has received a lot of interest. Du *et al*. (2023) proposed a brain-visual-linguistic representation learning framework, called BraVL, for generic neural decoding. BraVL leverages the mixture-of-product-of-experts (MoPoE) formulation (Sutter, Daunhawer, and Vogt 2021) and mutual information (MI) maximization regularization to learn consistent joint representation.

Zero-Shot Transfer Approaches. Classical zero-shot learning aims to deal with a prediction problem in unseen classes via transferring the knowledge learned from the seen classes. Early works (Hubert Tsai, Huang, and Salakhutdinov 2017; Han et al. 2021; Schonfeld et al. 2019a) perform zero-shot classification by mapping between visual representations and the pre-trained word embedding of class names or attributes. Recently, large-scale image-text pre-training, represented by CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), is encouraging, which provides a new light on the zero-shot transfer. They show promising results on various downstream tasks, *e.g.*, image captioning (Mokady, Hertz, and Bermano 2021), and image-text retrieval (Baldrati et al. 2022). The virtue of zero-shot transfer is that it avoids fine-tuning the model on the task-specific dataset (Yu, Seo, and Son 2023). The BrainSem, following the zero-shot transfer paradigm, is designed to solve the generic neural representation learning problem, which may lead to interesting applications.

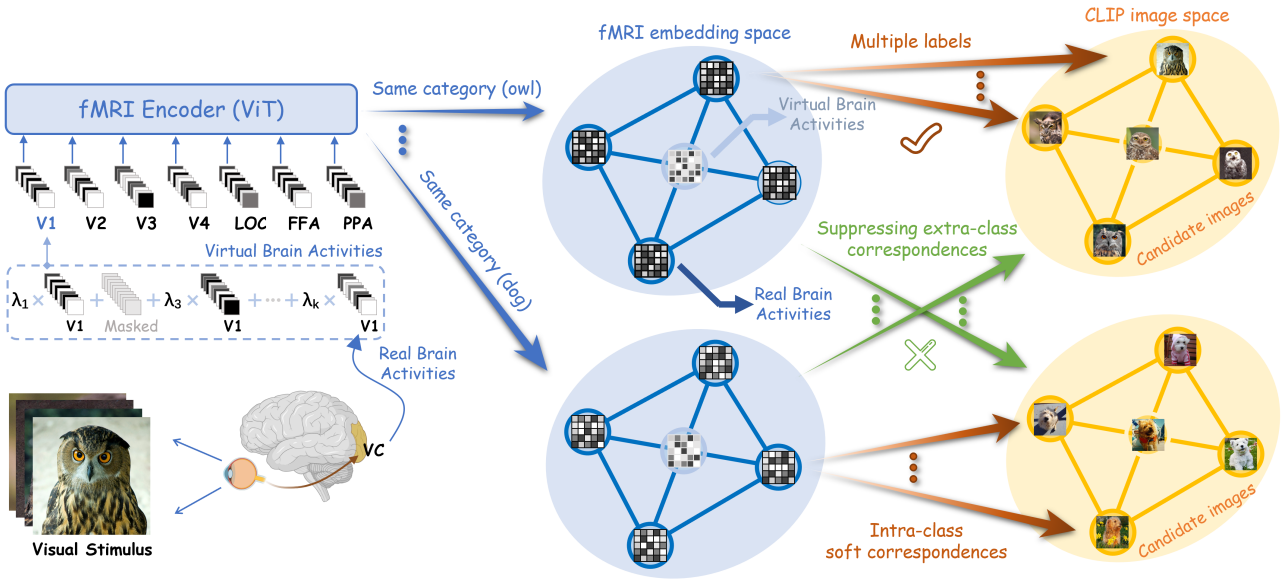

Figure 2: The overall pipeline of our BrainSem, including a fMRI encoder, and an image encoder. Our matching strategy is straightforward: The virtual brain activity generated within the same base vectors can be inferred to represent any instance of the same class, as they share identical category semantic properties, which not only guarantees matching flexibility but also conforms to local consensus principle.

Approach

Fig. 2 shows the overall pipeline of the proposed BrainSem, which affords great cognitive and associative flexibility, thus allowing inference about never-experienced stimuli.

Dataset. In this work, we leverage a popular benchmark fMRI-image dataset, termed as GOD (Horikawa and Kamitani 2017). This dataset provides the brain fMRI signals of five subjects, which are recorded using a 3.0 Tesla Siemens MAGNETOM Trio scanner (TR, 3 s, voxel size, $3\times3\times3\ \mathrm{mm}^3$), while they are viewing natural images selected from ImageNet (Deng et al. 2009). Briefly, the GOD comprises a total of 1250 stimulus images drawn from 200 ImageNet categories, where 1200 images from 150 categories (eight from each category) are used in training sessions and the remaining are the test images (one from each category). Note that images or categories have no overlap between the training and test phases. The fMRI signals underwent 3-D motion correction by SPM5, and then were co-registered to the anatomical image of the same slices and whole-head anatomical image. The standard retinotopy mapping procedures are used for the delineating of early visual areas, like V1, V2, V3, and V4. The higher visual cortex includes fusiform face area (FFA), lateral occipital complex (LOC), and parahippocampal place area (PPA), which are identified by conventional functional localizers. Please visit the original paper for more details of preprocessing.

Neuro-Semantic Re-Organization Format

The lack in labeled data place severe limitations on the generalization power of todays neural decoders, and individual variability further complicates this problem. From a neural encoding perspective, recent advances find that cognitive space is navigable, namely, each stimulus is located according to its attribute values, thus enabling the property or concept representations as convex regions of cognitive space (Bellmund et al. 2018). Therefore, similar to such spatial representational format, we suggest that the limited recorded brain activities should be considered as a set of properties to span a virtual neural response space describing a specific concept, thus creating virtual asymmetric correspondences for relaxing stimulus-response connection.

Spatial cognition navigation needs to determine attribute dimensions, but it is difficult to decouple property components from fMRI. A good approximation is to directly treat fMRIs that evoked by similar stimuli as the concept attributes (*i.e.*, basis vector). Then, a concept can be defined as a linear combination of prototype vectors:

$$\mathcal{X}^c = f\left(\{x_i^c\}_{i=1}^{\mathcal{K}}\right) = \lambda_1 x_1^c + I_1\lambda_2 x_2^c + \cdots + I_{\mathcal{K}-1}\lambda_{\mathcal{K}} x_{\mathcal{K}}^c,$$
$$\lambda_i = \frac{\exp(\alpha_i)}{\sum_{k=1}^{\mathcal{K}}\exp(\alpha_k)}, \tag{1}$$

where $\{x_i^c\}_{i=1}^{\mathcal{K}}$ denotes the available fMRI signals, c represents the type of visual stimulus such as owl, i.i.d. random variables $\{\alpha_i\}_{i=1}^{\mathcal{K}} \sim U(-1,1)$, and $\{I_i\}_{i=1}^{\mathcal{K}-1}$ obey the Bernoulli distribution with a parameter $p = 0.5$, which controls the number of base vectors. Intuitively, in Eq. 1, a property described by a fMRI signal constitutes the simplest form of a concept, and a new virtual fMRI signal $\mathcal{X}^c$ can be sampled in a convex region.

Fuzzy Visual-Neural Matching Paradigm

Given a set of image-fMRI pairs $z^c = \{(x_k^c, y_k^c)\}_{k=1}^{\mathcal{K}}$, where y_k^c denotes the corresponding visual stimulus of brain signal

x_k^c, we can use the proposed neuro-semantic re-organization format to create a lot of virtual brain activities, however, which may bring data imbalance problems due to limited stimulus images, and force the network to memorize specific visual features representing the image content. Adopting the mixup technique (Zhang et al. 2018), utilizing the same interpolation augmentation for labels, appears to be a plausible solution to this problem, but interpolation of labels does not introduce richer high-level visual semantic information. In fact, the reasonable fine-grained semantic information of virtual brain activities cannot be determined. So, we propose building semantic library for each concept and adopting a fuzzy strategy to relax the matching relationship, thus avoiding in-depth understanding of visual appearance details.

Specifically, we simply use the candidate images of the same class, which are selected from ImageNet (about 1k for each category), as the semantic library. Note that more complex correlations between visual stimuli and candidates are worth exploring, but we leave it to future work. The fuzzy matching strategy is also just as easy to work with, *i.e.*, multi-target label, which is determined by the number of images under the same category in a batch. Suppose $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^N$ is a randomly sampled minibatch with size N, where $\mathcal{X}_i$ is a virtual brain signal and $\mathcal{Y}_i$ is a candidate image, the labels can be defined as:

$$\left[\mathcal{J}\left(\{(\mathcal{X}_n, \mathcal{Y}_n)\}_{n=1}^N\right)\right]_{ij} = \begin{cases} 1, & \text{if } C(\mathcal{X}_i) = C(\mathcal{Y}_j), \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where $C(\cdot)$ returns the corresponding class label (1 donotes positive example, and 0 is negative example). Finally, our optimization target of fuzzy matching can be formulized as cross entropy across all positive pairs:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\left[\mathcal{J}\left(\{(\mathcal{X}_n, \mathcal{Y}_n)\}_{n=1}^N\right)\right]_{ij}\ell_{i,j}}{\sum_{n=1}^{N}\left[\mathcal{J}\left(\{(\mathcal{X}_n, \mathcal{Y}_n)\}_{n=1}^N\right)\right]_{in}}, \quad (3)$$

where the denominator term is a normalized factor for smoothing label (if there are more positive samples in a minibatch, the label will be smoother). Note that $\ell_{i,j}$ is the loss term based on pairwise cosine similarity, *i.e.*,

$$\ell_{i,j} = log\frac{\exp\left(f_b(\mathcal{X}_i)^{\mathrm{T}} f_v(\mathcal{Y}_j)/||f_b(\mathcal{X}_i)||_2||f_v(\mathcal{Y}_j)||_2\right)}{\sum_{n=1}^{N}\exp\left(f_b(\mathcal{X}_i)^{\mathrm{T}} f_v(\mathcal{Y}_n)/||f_b(\mathcal{X}_i)||_2||f_v(\mathcal{Y}_n)||_2\right)},$$

where $f_b(\cdot)$ is a fMRI embedding encoder, and $f_v(\cdot)$ is an image encoder to extract visual semantics representations. Two encoders can take any form, such as a Transformer (Vaswani et al. 2017) or a convolutional network (CNN), and we will give instances in the next section.

From a different perspective, the intuition behind of our fuzzy matching strategy is biologically plausible. We assume that fMRI signals, recording the variations in Blood Oxygen Level-Dependent (BOLD), not only preserves the information about the visual properties of stimulus, but also couples the context knowledge of what's relevant to seem visual object due to the spontaneous imagination function of the human brain. For example, when we see a triangular-shaped building, an Egyptian pyramid probably comes to mind. Several advanced studies (Chen, Qi, and Pan 2023; Du et al. 2023) have demonstrated that exploiting richer set of semantic features can help improve the decoding performance. This promising practice is naturally combined with our matching strategy, where multi-objective optimization provides more flexibility for mining potential contextual information, and negative samples across different classes makes the network prone to learn a significant spatial margin for fMRIs evoked by dissimilar stimuli.

Network Architecture

In the following, we will detail our specific implementation.

Image Encoder. Dual-coding theory considers that the encoding principles of concrete cognitive concepts are both visual and linguistical (Du et al. 2023). On the other hand, the visual representations of CLIP model (Radford et al. 2021), aligned with large-scale natural language features, also exhibit locality-sensitive properties. Therefore, the pre-trained CLIP embedding space would be an ideal target domain, and we use an image encoder of CLIP as a simplified visual-linguistic proxy to guide a fMRI encoder.

fMRI Encoder. Recent works have found that the performance of neural decoding can receive benefit in the divide and conquer decoding strategy (Fang, Qi, and Pan 2020; Takagi and Nishimoto 2023). The main reason behind is the hierarchical encoding properties in brain cortex. We also follow this practice, but go one step further and consider the interactions among V1-V4, LOC, FFA and PPA, which is more fine-grained than the division of early and higher visual areas. Specifically, a fMRI recording is split into 7 fixed-size patches based on the region of interests (ROIs) provided by GOD, and fed to a standard ViT. Note that the extra learnable classification token is regarded as the final brain embedding (see Fig.2 left).

Implementation Details. We use ViT (Dosovitskiy et al. 2021) as the backbone for the fMRI encoder due strong to the long-dependencies modeling ability. Specifically, the full model settings, similar to ViT-Large (Dosovitskiy et al. 2021), are as follows: patch size of 7 (number of brain areas), Transfomer encoder depth of 24, embedding dimension of 512, MLP size of 2048, heads number of 16. For image encoder, we leverage the pre-trained CLIP image encoder (ViT-B/16), which is frozen in the training phase, to guide fMRI embedding. Different architecture choices are explored in our ablation study. The size of the candidate set (including 150 categories) is 200.7k, which is collected from ImageNet (Deng et al. 2009). Note that the classes of the candidate images do not overlap with the test set.

In the training phase, our batch size is 200, and Adam solver (Kingma and Ba 2014) is used for the optimization of the fMRI encoder parameters, with a learning rate of 1e-4. Since the number of fMRI voxels varies with different brain visual areas. Therefore, to facilitate data processing, we use zero to fill them to a uniform size before feeding the fMRI encoder. The BrainSem (implemented by Pytorch) pre-training is performed on 4 NVIDIA GeForce RTX3090 GPUs until the model converges.

Method	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		**Average**	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CADA-VAE (CVPR19) (Schonfeld et al. 2019b)	6.3%	35.7%	6.5%	40.1%	17.7%	54.3%	12.2%	36.6%	7.5%	35.0%	10.0%	40.3%
MVAE (NIPS18) (Wu and Goodman 2018)	5.8%	31.5%	5.4%	38.5%	17.1%	52.5%	14.0%	40.9%	7.9%	34.6%	10.0%	39.6%
MMVAE (NIPS19) (Shi et al. 2019)	6.6%	38.7%	6.6%	41.0%	22.1%	56.3%	14.5%	42.5%	8.5%	38.1%	11.7%	43.3%
MoPoE-VAE (ICLR21) (Sutter, Daunhawer, and Vogt 2021)	8.5%	44.0%	8.3%	48.1%	22.7%	61.8%	14.6%	58.5%	10.5%	46.4%	12.9%	51.8%
BraVL (TPAMI23) (Du et al. 2023)	9.1%	46.8%	8.9%	**48.9%**	24.0%	**62.1%**	15.1%	60.0%	12.9%	47.9%	14.0%	53.1%
Ours	**24.0%**	**50.0%**	**24.0%**	44.0%	**28.0%**	58.0%	**26.0%**	**62.0%**	**26.0%**	**54.0%**	**25.6%**	**53.6%**

Table 1: Zero-shot classification accuracy of several neural decoding methods. All compared approaches are trained with the visual (V), textual (T), as well as combined (V&T) features, and the results are taken from BraVL. The best is printed in Bold.

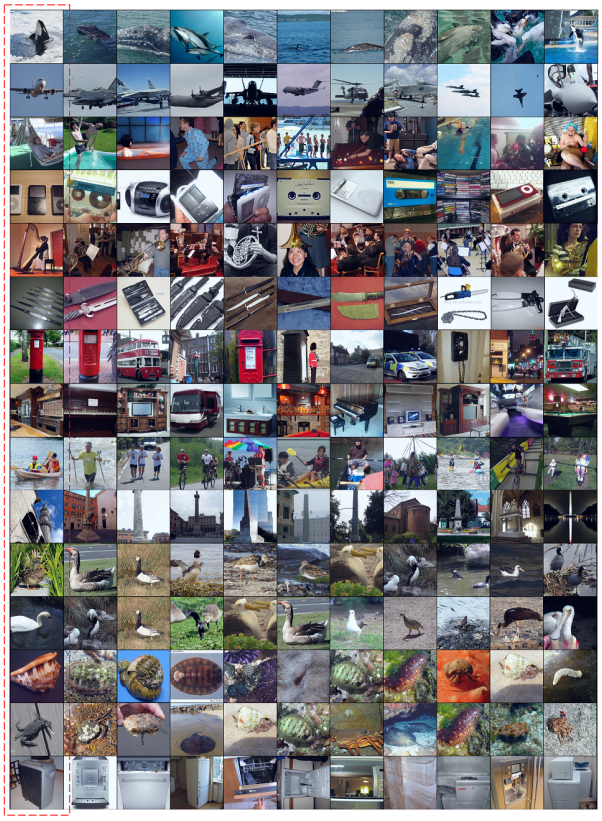

Figure 3: Image retrieval results of BrainSem on a random subset (involving 10k images) of ImageNet-1k. From left to right: the visual stimuli, and Top-10 most similar images searched by fMRI representation. Note that, the categories of the test visual stimuli were not included in the visual-neural representation learning, thus it is a zero-shot process.

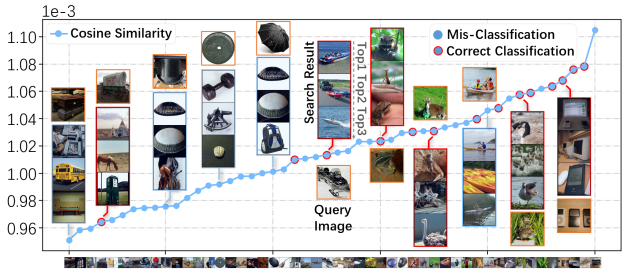

Figure 4: The relationship between decoding accuracy and the inherent similarity of visual stimuli about GOD training/test split. Each red circle represents a successful decoding case, and the corresponding visual stimulus is plotted on the horizontal axis. Note that the Top-3 images are retrieved in the training set by conditioning on test images.

Experimental Results

Evaluation with Brain-Image Retrieval

To evaluate the quality of learned neural semantic representation, we test BrainSem with the near-duplicate retrieval problem, to see the semantic content of neural representations by their nearest images retrieved in the representation space. Formally, given a query fMRI signal, our target is to retrieve the images that are most similar to the corresponding visual stimulus from a database and return a ranked list. To this end, we first calculate the similarities between the representation of query fMRI generated by BrainSem and all the CLIP image embeddings, and then sort the similarities to return a ranked list. Note that, for all the visual stimuli used in this experiment, their categories were not included in the visual-neural representation learning, thus it is zero-shot.

The fMRI-image search results from a random subset (10k) of ImageNet-1k are shown in Fig. 3. From the retrieval results, we see that the representations with BrainSem retrieve images with similar semantic content mostly. In the first example where the stimulus picture was a whale, the top retrievals are mostly whales. Similar results can be found in most examples. It is interesting to find that, for some cases, although the top retrievals fall into the wrong categories, they still show strong semantic correlations. For example, in the case of 'iPod' (the 4^{th} row), the wrong retrievals include tapes and recorders, which are also associated with music play. Another interesting case is the 'postbox' (the 7^{th} row), some of the wrong retrievals show 'London' elements, which may reflect the associations of subject. In the case of 'boating' (the 9^{th} row), the wrong retrievals are mostly related to sport, which is semantically correlated with the stimulus. The result demonstrates that the visual and neural semantic representations are well-matched, and the learned representation is transferable to unknown categories in a zero-shot way. It also indicates BrainSem gains a

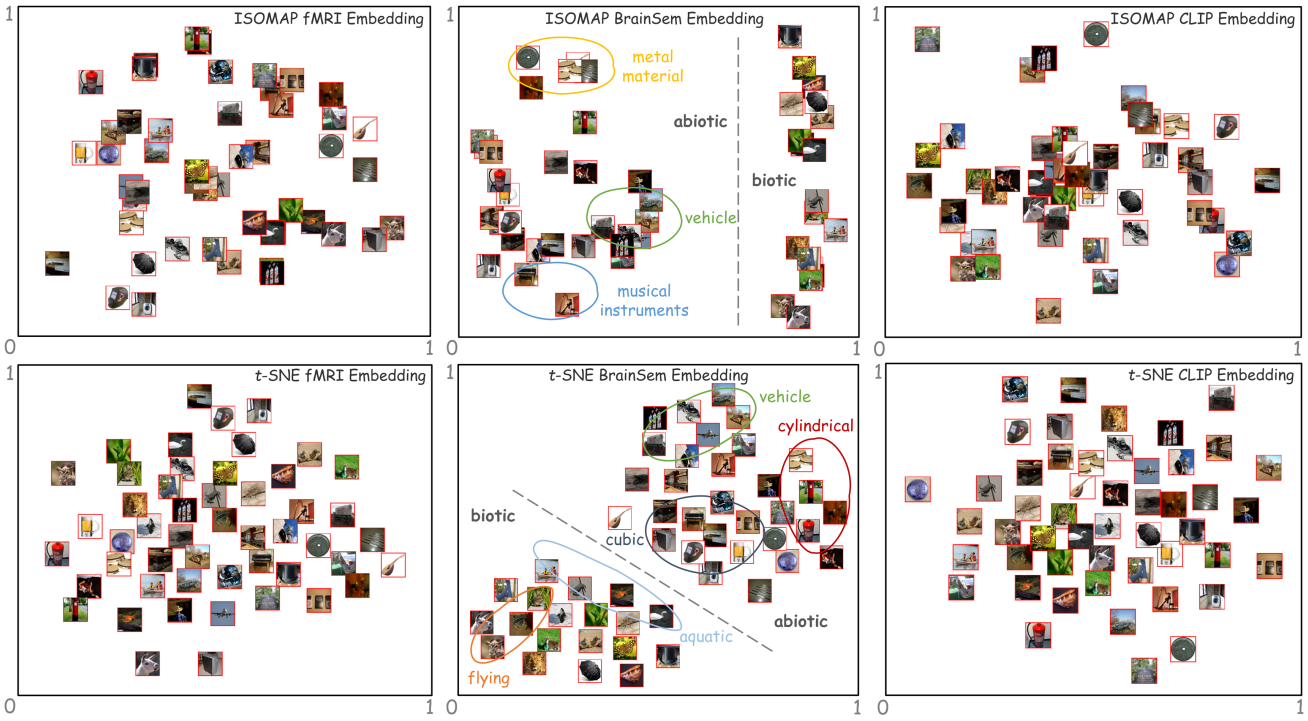

Figure 5: ISOMAP and t-SNE embedding visualization. From left to right: raw fMRI representations, neural representations learned from our BrainSem, pre-trained CLIP visual representations.

generalizable semantic understanding of data.

Analysis of the Neural Semantic Representations

To understand how BrainSem can retrieve in a zero-shot way, we investigate the relationship between 'good' and 'bad' cases. Specifically, we analyze the visual and semantic similarity between test set and training set by calculating the cosine similarity between CLIP image features. The results are presented in Fig. 4. We find misclassification tends to primarily occur when the sample has a large semantic gap with the known classes, which is in line with our intuition.

Then, in order to further understand the learned neural semantic representation quality, we visualize representation distributions of BrainSem by projecting them into two-dimensional plane with both ISOMAP (Balasubramanian and Schwartz 2002), and t-SNE (Van der Maaten and Hinton 2008), as shown in Fig. 5 center. From visualization results, we observe that the embedding space of BrainSem emerges with similar properties to cognitive spaces that defined by feature dimensions satisfying geometric constraints, namely, stimuli can be located based on their feature values along the dimensions (Bellmund et al. 2018). In particular, *the representation distribution of BrainSem shows distinct biotic and abiotic regions via a clear linear boundary*.

It is not yet clear whether this phenomenon is intrinsic to the fMRI signal or guided by CLIP's supervision. For exploring the mechanism, we also put the corresponding raw fMRI responses and CLIP features at the low-dimensional space, as shown in Fig. 5 left and right, respectively. It is interesting to find that, compared to our method, the fMRI data and CLIP visual representations demonstrate weak locality-sensitive property (w/o a clear boundary between the living and non-living categories) in both ISOMAP and t-SNE.

Generally, the locality-sensitive encoding of concepts lays the foundation for the understanding of how our brain store and retrieve information effectively and robustly with large-scale data. In DL era, semantic representations with locality-sensitive structure learned with neural networks is by no means new. However, good representation usually relies on high computing power and large-scale datasets, which is very inefficient and in stark contrast to the human brain.

The results suggest that our method learns biological-valid locality-sensitive structure from limited data, which support the findings in (Dasgupta, Stevens, and Navlakha 2017; Bellmund et al. 2018), so that the neural representation distributions for fMRI contains a more clear clusters representing different properties or concepts. BrainSem may offer hope to improve generalization for learning on limited data, and provide a kit for processing spatial navigation involving complex, multidimensional concepts, which is a vital problem in human cognition (Bellmund et al. 2018).

Comparison with Zero-Shot Classification

Here we evaluate the neural semantic representations with a zero-shot classification task and compare with existing approaches. In classification problems, zero-shot usually aims for seeking generalizing to unseen categories, which is a good way of measuring the task learning capability of a

Encoder	ViT-L/14	ViT-B/16	ViT-B/32	ResNet50
Top-1 Acc	**25.8%**	25.6%	25.2%	24.8%
Top-5 Acc	**54.4%**	53.6%	52.0%	51.8%

Table 2: Ablation tests on pre-trained CLIP image encoder. Note that all parameters are set to the default (except for the image encoder). The best is printed in Bold.

model (Radford et al. 2021). To our best knowledge, BraVL is the current SOTA in this field, which first studied multimodal learning of brain-visual-linguistic features for zero-shot neural decoding. Our method is learning to predict if an fMRI recording and an image are paired semantically, which is more accurate while being simpler, in contrast to the generative model BraVL that optimization based on maximum MI. The zero-shot transfer performances are summarized in Tab. 1. The well-trained BrainSem model obtains an optimal performance both in Top-1 and Top-5 accuracy on GOD, where Top-1 accuracy is 82.9% higher than the SOTA. Note that decoding at the group level remains a problem difficult to work out due to individual variability (Horikawa and Kamitani 2017; Beliy et al. 2019; Chen et al. 2023b), the training and test are performed on identical subject.

Model Evaluation and Ablation Studies

Architecture Configurations. To systematically study the impact of architecture settings, we repeat the neural decoding experiments with BrainSem, but this time focusing on several architecture variants. First of all, to reveal the effect of CLIP image encoder on the performance of our method, we consider four different pre-trained CLIP models, including ViT-L/14, ViT-B/16, ViT-B/32, and ResNet50, which are summarized in Tab. 2. The results proved that larger CLIP model may improve the performance, albeit modestly. Specifically, ViT-L/14 model achieves a lead of 0.2% to 1% on the Top-1 accuracy. Second, we leverage ViT-B/16 as the default model, which requires substantially less computational resources, to evaluate the influence of the size of fMRI encoder on decoding performance, as reported in Tab. 3. A few patterns can be found here. In the case of using the fuzzy matching paradigm, overall decoding performance is barely affected by the fMRI encoder architecture choice. When training only on the raw annotation data, the smaller model outperforms marginally larger model.

Matching Strategy. To understand the effects of the proposed fuzzy visual-neural matching strategy, and the importance of augmentation semantic library. We mainly demonstrate the performance of BrainSem when applying candidate set and virtual fMRI signals individually. The ablation study results are presented in Tab. 3. Clearly, the virtual asymmetric correspondences created by neuro-semantic reorganization format bring a substantial performance boost to the decoding accuracy. For example, in the last configuration, using simultaneously virtual brain responses and candidate set improves Top-1 accuracy from 18.8% to 25.6%. Individual data augmentations also achieve 1.6% and 5.2% relative improvements, respectively.

fMRI Encoder		Image Encoder	Batch Size	M.		Mean Acc.	
Layer	Heads			V.	B.	Top-1	Top-5
6	4	ViT-B/16	100	×	×	19.4%	45.8%
6	4	ViT-B/16	100	✓	×	21.0%	48.2%
6	4	ViT-B/16	100	×	✓	24.2%	**54.4%**
6	4	ViT-B/16	100	✓	✓	**25.4%**	52.8%
6	4	ViT-B/16	200	×	×	19.4%	46.4%
6	4	ViT-B/16	200	✓	×	20.0%	50.6%
6	4	ViT-B/16	200	×	✓	23.8%	**55.2%**
6	4	ViT-B/16	200	✓	✓	**25.2%**	53.2%
12	8	ViT-B/16	100	×	×	19.0%	46.2%
12	8	ViT-B/16	100	✓	×	20.6%	49.0%
12	8	ViT-B/16	100	×	✓	24.0%	54.4%
12	8	ViT-B/16	100	✓	✓	**24.8%**	**58.4%**
12	8	ViT-B/16	200	×	×	19.2%	46.6%
12	8	ViT-B/16	200	✓	×	21.0%	48.8%
12	8	ViT-B/16	200	×	✓	23.8%	54.6%
12	8	ViT-B/16	200	✓	✓	**25.2%**	**59.2%**
24	16	ViT-B/16	100	×	×	18.2%	43.2%
24	16	ViT-B/16	100	✓	×	21.2%	50.0%
24	16	ViT-B/16	100	×	✓	23.6%	**54.0%**
24	16	ViT-B/16	100	✓	✓	**25.4%**	53.2%
24	16	ViT-B/16	200	×	×	18.8%	44.0%
24	16	ViT-B/16	200	✓	×	20.4%	49.2%
24	16	ViT-B/16	200	×	✓	24.0%	**54.0%**
24	16	ViT-B/16	200	✓	✓	**25.6%**	53.6%

Table 3: Ablation tests on fMRI encoder, batch size, and fuzzy visual-neural matching. M. denotes matching strategy, where V. stands for visual semantic library, and B. represents virtual brain activity.

Conclusion

We propose a simple, yet effective neural fuzzy matching network, called BrainSem, to explore the locality-sensitive homologies between neural and artificial representation latent spaces. The major challenge here is how to extract high-level semantic components from limited neural signals, given that the human perceptual-cognitive system is inherently multi-modal. The proposed fuzzy (one-to-many) matching with specially designed data augmentation is the key technical component in finding an alignment between the cross-modal representations. That is, due to the multi-objective supervision signals, this one-to-many matching strategy encourages the models to capture shared category-level semantics representations, which results in a nontrivial and meaningful neurosemantic contrastive learning framework. Our findings claimed that, using only a small set of fMRI recordings for semantic space alignment, we could obtain a valid neural embedding for zero-shot classes. Therefore, we think our BrainSem provided potential evidence for the structure similarity in the semantic embedding space between the brain and large artificial model (CLIP), and suggested that similarity-based learning could be the key rule shared between the brain and machine. Overall, we hope the proposed neural semantic matching framework will inspire new intelligence paradigms and provide a tool for cognitive neuroscience.

Acknowledgments

This work was supported in part by the Science and Technology Innovation 2030 Major Projects (2021ZD0200400), the Key Research and Development Program of Zhejiang Province in China (2020C03004), and the Natural Science Foundation of China (NSFC) (Nos. 61925603, U1909202, and 62276228).

References

Ahmed, H.; Wilbur, R. B.; Bharadwaj, H. M.; and Siskind, J. M. 2021. Object classification from randomized EEG trials. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3845–3854.

Akamatsu, Y.; Harakawa, R.; Ogawa, T.; and Haseyama, M. 2020. Brain decoding of viewed image categories via semi-supervised multi-view Bayesian generative model. *IEEE Transactions on Signal Processing*, 68: 5769–5781.

Balasubramanian, M.; and Schwartz, E. L. 2002. The isomap algorithm and topological stability. *Science*, 295(5552): 7–7.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022. Effective conditioned and composed image retrieval combining CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21466–21474.

Beliy, R.; Gaziv, G.; Hoogi, A.; Strappini, F.; Golan, T.; and Irani, M. 2019. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. *Advances in Neural Information Processing Systems*, 32.

Bellmund, J. L. S.; Grdenfors, P.; Moser, E. I.; and Doeller, C. F. 2018. Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415): eaat6766.

Chen, J.; Qi, Y.; and Pan, G. 2023. Rethinking Visual Reconstruction: Experience-Based Content Completion Guided by Visual Cues. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 4856–4866. PMLR.

Chen, J.; Qi, Y.; Wang, Y.; and Pan, G. 2023a. MindGPT: Interpreting What You See with Non-invasive Brain Recordings. *arXiv preprint arXiv:2309.15729*.

Chen, Z.; Qing, J.; Xiang, T.; Yue, W. L.; and Zhou, J. H. 2023b. 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*, 22710–22720.

Damarla, S. R.; and Just, M. A. 2013. Decoding the representation of numerical values from brain activation patterns. *Human Brain Mapping*, 34(10): 2624–2634.

Dasgupta, S.; Stevens, C. F.; and Navlakha, S. 2017. A neural algorithm for a fundamental computing problem. *Science*, 358(6364): 793–796.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Du, C.; Du, C.; Huang, L.; Wang, H.; and He, H. 2022. Structured Neural Decoding With Multitask Transfer Learning of Deep Neural Network Representations. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2): 600–614.

Du, C.; Fu, K.; Li, J.; and He, H. 2023. Decoding Visual Neural Representations by Multimodal Learning of Brain-Visual-Linguistic Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–17.

Fang, T.; Qi, Y.; and Pan, G. 2020. Reconstructing perceptive images from brain activity by shape-semantic GAN. *Advances in Neural Information Processing Systems*, 33: 13038–13048.

Fang, T.; Zheng, Q.; and Pan, G. 2023. Alleviating the Semantic Gap for Generalized fMRI-to-Image Reconstruction. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gaziv, G.; Beliy, R.; Granot, N.; Hoogi, A.; Strappini, F.; Golan, T.; and Irani, M. 2022. Self-supervised Natural Image Reconstruction and Large-scale Semantic Classification from Brain Activity. *NeuroImage*, 254: 119121.

Han, Z.; Fu, Z.; Chen, S.; and Yang, J. 2021. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2371–2381.

Haxby, J. V.; Gobbini, M. I.; Furey, M. L.; Ishai, A.; Schouten, J. L.; and Pietrini, P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539): 2425–2430.

Horikawa, T.; and Kamitani, Y. 2017. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1): 1–15.

Hubert Tsai, Y.-H.; Huang, L.-K.; and Salakhutdinov, R. 2017. Learning robust visual-semantic embeddings. In *Proceedings of the IEEE International conference on Computer Vision*, 3571–3580.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Kamitani, Y.; and Tong, F. 2005. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5): 679–685.

Kay, K. N.; Naselaris, T.; Prenger, R. J.; and Gallant, J. L. 2008. Identifying natural images from human brain activity. *Nature*, 452(7185): 352–355.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Palazzo, S.; Spampinato, C.; Kavasidis, I.; Giordano, D.; Schmidt, J.; and Shah, M. 2020. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11): 3833–3849.

Popham, S. F.; Huth, A. G.; Bilenko, N. Y.; Deniz, F.; Gao, J. S.; Nunez-Elizalde, A. O.; and Gallant, J. L. 2021. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11): 1628–1636.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019a. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019b. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Shi, Y.; Paige, B.; Torr, P.; et al. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32.

Spampinato, C.; Palazzo, S.; Kavasidis, I.; Giordano, D.; Souly, N.; and Shah, M. 2017. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6809–6817.

Stevens, C. F. 2015. What the flys nose tells the flys brain. *Proceedings of the National Academy of Sciences*, 112(30): 9460–9465.

Sutter, T. M.; Daunhawer, I.; and Vogt, J. E. 2021. Generalized Multimodal ELBO. In *International Conference on Learning Representations*.

Takagi, Y.; and Nishimoto, S. 2023. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14453–14463.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Van Gerven, M. A.; Cseke, B.; De Lange, F. P.; and Heskes, T. 2010. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*, 50(1): 150–161.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wu, M.; and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31.

Wu, Z.; Pan, G.; and Zheng, N. 2013. Cyborg Intelligence. *IEEE Intelligent Systems*, 28(5): 31–33.

Yamins, D. L.; and DiCarlo, J. J. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3): 356–365.

Yargholi, E.; and Hossein-Zadeh, G.-A. 2016. Brain decoding-classification of hand written digits from fMRI data employing Bayesian networks. *Frontiers in human neuroscience*, 10: 351.

Yu, S.; Seo, P. H.; and Son, J. 2023. Zero-shot Referring Image Segmentation with Global-Local Context Features. *arXiv preprint arXiv:2303.17811*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.