NeuroVLM: A generative vision-language framework for human neuroimaging

Ryan Hammonds[2*], Jerjes Aguirre-Chavez[2], Borngreat Omoma-Edosa[2], Bradley Voytek[1,2,3,4]

[1*]Department of Cognitive Science, UC San Diego, La Jolla, CA, USA.
[2]Halıcıoğlu Data Science Institute, UC San Diego, La Jolla, CA, USA.
[3]Neurosciences Graduate Program, Organization, La Jolla, CA, USA.
[4]Kavli Institute for Brain and Mind, Organization, La Jolla, CA, USA.

*Corresponding author(s). E-mail(s): rphammonds@ucsd.edu;
Contributing authors: jaguirrechavez@ucsd.edu;
oomomaedosa@ucsd.edu; bvoytek@ucsd.edu;

Abstract

Neuroimaging research has produced tens-of-thousands of articles that pair natural language and activation coordinate tables. Recent advances in vision-language models (VLMs) have provided methods to model text and images simultaneously. In this work, we present NeuroVLM, a model architecture for learning from 30,000 human neuroimage-text pairs. The architecture supports contrastive and generative objectives. The contrastive model ranks similarity between neuroimages and text. The generative models include text-to-neuroimage and neuroimage-to-text. These models are evaluated on network images from a variety of atlases, statistical maps from diverse publications, and images created from coordinate tables. These models are capable of generating atlases or maps given a text corpus, generating text interpretations of neuroimages, labeling networks, finding publications most related to a neuroimage query, or finding neuroimages most related to a text query.

Keywords: neuroimaging, vision-language models, generative artificial intelligence, fMRI

1 Introduction

Over the past three decades, human functional magnetic resonance imaging (fMRI) neuroimaging research has changed the landscape of our understanding of how the brain gives rise to behavior and cognition. Tens of thousands of experiments have been performed, resulting in a massive, multimodal corpus of peer-reviewed publications that link natural language text to spatial brain coordinates. While powerful, expert decisions on how to interpret these results and which new experiments to run are constrained by one's own degree of familiarity with the literature. These constraints result in, for example, variations in how individual brain regions are named, or how collections of brain regions are grouped together and named as common networks [1]. Inconsistencies in nomenclature and interpretations in fMRI are exacerbated by small effect sizes [2–4] and large variability in data analysis approaches. Even when working on the same datasets, different research groups can come to different conclusions [5]. Furthermore, many fMRI studies are interpreted in isolation from other human brain maps, such as gene expression, neurotransmitter and modulator distribution, and electrophysiology [6], giving rise to a narrow view of the brain-behavior relationship [7].

Prior to the introduction of modern large language models (LLMs), text-mining of neuroscientific papers was demonstrated to be capable of generating plausible hypotheses by identifying gaps between text concepts mapped to a network graph [8]. In other domains, such as materials science, high-dimensional text embeddings have been shown to be capable of recapitulating concepts such as the periodic table of elements, and complex structure–property relationships in materials [9].

To overcome some of these limitations, a large, diverse training corpus is useful for integrating information across a broad neuroscientific context [7, 10]. This includes diverse datasets that pair text and brain activation maps, enabling models to translate between modalities: text-to-neuroimage and neuroimage-to-text. Past text-to-neuroimage models such as neurosynth [3] and neuroquery [11] use term-frequencies based on a predefined corpus [3, 11]. This limits handling of unseen phrases, synonymy, and context, and assumes near-independence of terms or linear additivity across terms. These methods fail to generalize to out-of-corpus language and to account for non-linear interactions.

Advances in language modeling architectures [12–14] move beyond term-frequencies and instead use transformers to model language. Transformer architectures are standard in LLMs, remove the dependence on a fixed vocabulary, and are context-sensitive. Transformers are also used in vision-language models (VLMs), with the goal of aligning text-image latent vector pairs [15].

Recent works such as neurocontext [16] and niclip [17] use LLMs to encode text from neuroimaging publications. Contrastive loss [15] was used in these studies to learn a bi-modal embedding space, aligning text and neuroimage representations. These models provide similarity-based ranking and retrieval, given a corpus of publications and neuroimages. While powerful, contrastive models cannot generate brain images from input text, nor text outputs from 3D brain maps.

Here we present, NeuroVLM (Fig. 1a), a lightweight, artificial intelligence (AI) framework that trains both generative and contrastive models. This framework embeds

text-neuroimage pairs to a shared latent space. This latent space can be trained with a contrastive objective to support ranking and retrieval, which allows us to label neuroimages, e.g., spatial ICA components, with text and select neuroimages most related to a text query. Decoding the latent space provides generative text-to-neuroimage and neuroimage-to-text models.

To improve the accessibility of these models, we significantly reduced the computational requirements by using much smaller text encoders [18] with 110 million parameters, rather than depending on LLMs with tens to hundreds of billions of parameters. Each component of NeuroVLM is modular and interchangeable with larger or smaller models depending on requirements.

The resulting open-source models support a variety of applications. Use cases include generating atlases or maps from a text corpus or query, generating natural-language interpretations of human brain maps (Fig. 2), labeling functional networks (Fig. 4b), and supporting cross-modal retrieval by finding the publications most related to a statistical map or the statistical maps most related to a text query.

2 Methods

2.1 Datasets

Data were aggregated from PubMed central, neurosynth compose [19], cognitive atlas [20], the network correspondence toolbox [1], and NeuroVault [21]. The networks toolbox maps [1], NeuroVault maps [21], and a subset of the PubMed articles were left out as test sets. Cogatlas is a dataset of text labels used to label activation maps post-training.

1. PubMed Central: 30,000 publications with titles, abstracts, and coordinate tables
2. Cognitive Atlas [20]: 1,000 cognitive terms and definitions
3. Networks [1]: 100 labeled network images
4. NeuroVault [21]: 3,000 group-level activation maps, titles, and abstracts

2.2 Training

Training was performed with 30,826 text-neuroimage pairs, primarily from PubMed. Text included titles and abstracts of publications. Neuroimages are defined as sparsely clustered, MNI-space activation maps. PubMed training articles instead provide coordinate tables, these tables were smoothed in 3D, MNI space. Continuous statistical maps are provided in the network and neurovault datasets; these were used as additional test sets. DOIs used in the PubMed training set were excluded from all test sets.

2.3 Coordinate Smoothing

Coordinates were convolved in MNI space with a 9 mm activation likelihood estimation (ALE) kernel, following [11]. The resulting activation maps were transformed using the dictionaries of functional modes (DiFuMo) atlas [22]. Reconstruction replaces voxel values with a weighted combination of DiFuMo components. Component coefficients

were computed as the DiFuMo-weighted mean activation and were used to reconstruct voxel space by re-expanding the coefficients through the DiFuMo voxel maps. This returns a voxel-space map, $\hat{\mathbf{X}}$, that is a linear combination of DiFuMo components (see Sec. 6.1 for details).

2.4 Model Architecture

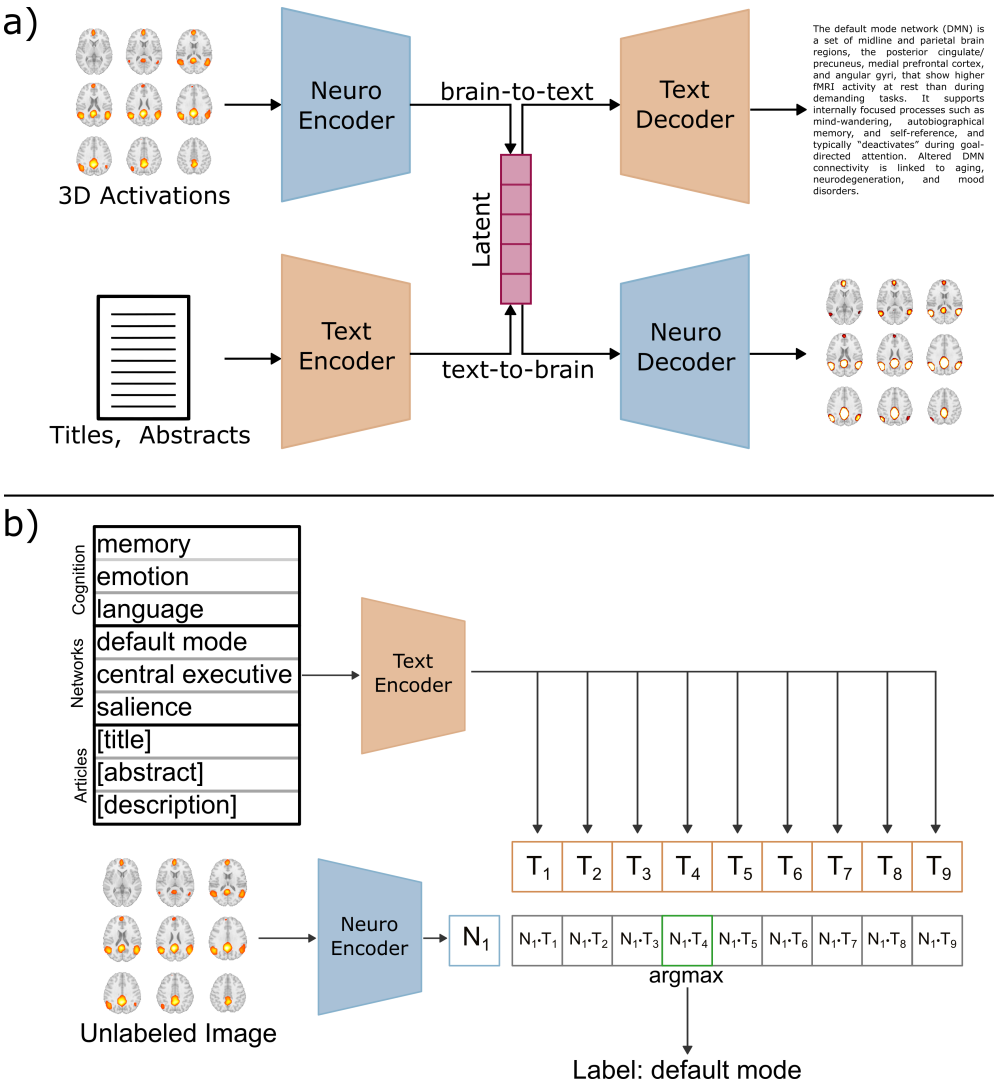


Fig. 1 Model architecture. a) NeuroVLM maps pairs of activation maps and text, from the same publication, to a shared latent vector (pink). Decoding the latent vector results in brain-to-text or text-to-brain generation. **b)** The shared latent vector is the basis for contrastive models, e.g., ranking and retrieving text most similar to a query brain map. Given an unlabeled, query activation map, the method ranks how similar a corpus of text is to the image. All encoded vectors, N_1 and $\{T_1, \ldots, T_9\}$ are unit normalized. The inner product between these vector pairs is cosine similarity. Taking argmax gives the most similar text. Any text corpus or image set may be used, suitable for a variety of use cases. The direction may also be flipped from what is presented: comparing a single text query to a set of activation maps.

Building upon past work, we outline a general vision-language framework for neuroimaging applications: NeuroVLM (Fig. 1a). The encoders in this framework map from high-dimensional text and activation-map space to low-dimensional latent vectors (Fig. 1a, pink). After training, pairs of text and neuroimages encode to highly similar vectors. Decoders are able to generate text or activation maps from latent vector embeddings. These encoders and decoders collectively provide text-to-image and image-to-text generation. This framework is general and does not make assumptions on the choice of encoder or decoder architectures.

2.5 Autoencoder

An autoencoder was trained to compress 3D activation maps to 368-dimensional latent vectors (Fig. 1a, blue). Maps were down-sampled, masked, flattened, and binarized before passing through fully-connected linear layers with rectified linear unit (ReLU) activations. The downsampling and masking procedure was adapted from [11] and moves from a 1M dimensional MNI space to 28,000 dimensional space. The encoder transforms from 28,000 to a 368 dimensional latent representation. A key component to successful training was the binarization of the activation map prior to training and the use of binary cross entropy loss. Any continuous map, e.g., a z-score or p-value map, can be binarized with an appropriate threshold and clustering method. Another key component in training was sparsity, reducing the influence of dense maps that were fully activated and lacked sparse clusters of activation.

2.6 Text Encoder

NeuroVLM uses Specter [18], a small, efficient, and accessible transformer. This small model allows deploying the model on edge devices with little computational requirements. Specter is based on SciBERT [23] and incorporates citation-based connectivity graphs. Specter encodes title-abstract pairs to 768-dimensional latent vectors. Despite being trained on title-abstract pairs, Specter has an ad hoc query adapter, suitable for short form user queries.

2.7 Contrastive Learning

Contrastive learning maximizes cosine similarity between true text-image pairs and minimizes similarity between mismatched pairs [15, 17]. After training, unlabeled images can be labeled by ranking cosine similarity between the target image and any text corpus (Fig. 1b). Pretrained encoders can be used to label neuroimages with networks [1], cognitive concepts [20], or publication descriptions. Inversely, a set of images may be ranked in similarity, relative to target text.

2.8 Generative Models

2.8.1 Brain-to-Text: Generating Interpretations

Network labeling provides a brain-to-text model given an *a priori* corpus of text. This text corpus may be tailored to the users interest: networks (e.g., sensorimotor or

somatomotor network), regions (e.g., primary motor cortex), cognition (e.g., voluntary movement execution), tasks (e.g., finger-tapping tasks), or phenotypes (e.g., Parkinson's disease). These concepts are classes ranked based on similarity to the input image. The top most similar set of concepts are passed to an LLM to generate a summary (Fig. S1a). Users' queries or instructions can be included to further steer LLM generation. The concepts may also be tailored to user preferences and expectations. As an alternative to using concept's predictions, the latent neuroimage embedding may be mapped directly to a token embedding matrix and directly decoded by the LLM [24].

2.8.2 Text-to-Brain: Generating Activation Maps

Activation maps are generated by projecting text embeddings to the latent space of the neuroimage autoencoder and using the neuroimage decoder (Fig. S1b). Specter2 [25] is used to embed arbitrary text: titles, abstracts, descriptions, or user queries. A small neural network, the projection head, maps these embeddings to the shared latent space. The projection head allows the language model weights to be frozen while providing fine-tuning, similar to low-rank adaptation (LoRA) adapters [26]. The neuroimage decoder then transforms the latent embeddings to MNI space.

2.9 Performance Metric

The performance metric used for test evaluation was the recall@k curve. This measure was used for both contrastive and generative models. Cosine similarity between all pairs of latent text and neuroimages is computed. Recall@k measures how often the true neuroimage-text pair was found in the top-k most similar pairs [16]. Given a reference latent neuroimage, recall@k measures how often the paired latent text is in the top-k most similar pairs. The inverse is performed using a reference latent text and computing recall@k against latent neuroimages. We step k from 1 to the max size of the dataset and compute recall@k for each threshold. The area under this curve is reported, where 0.5 is chance performance and 1.0 is perfect performance. Recall@k is used for the contrastive model, examining how well aligned the text and neuroimage representations are. For generative models, recall@k is computed from latent representations of generated outputs and true targets.

3 Results

3.1 Generative Models

3.1.1 Networks

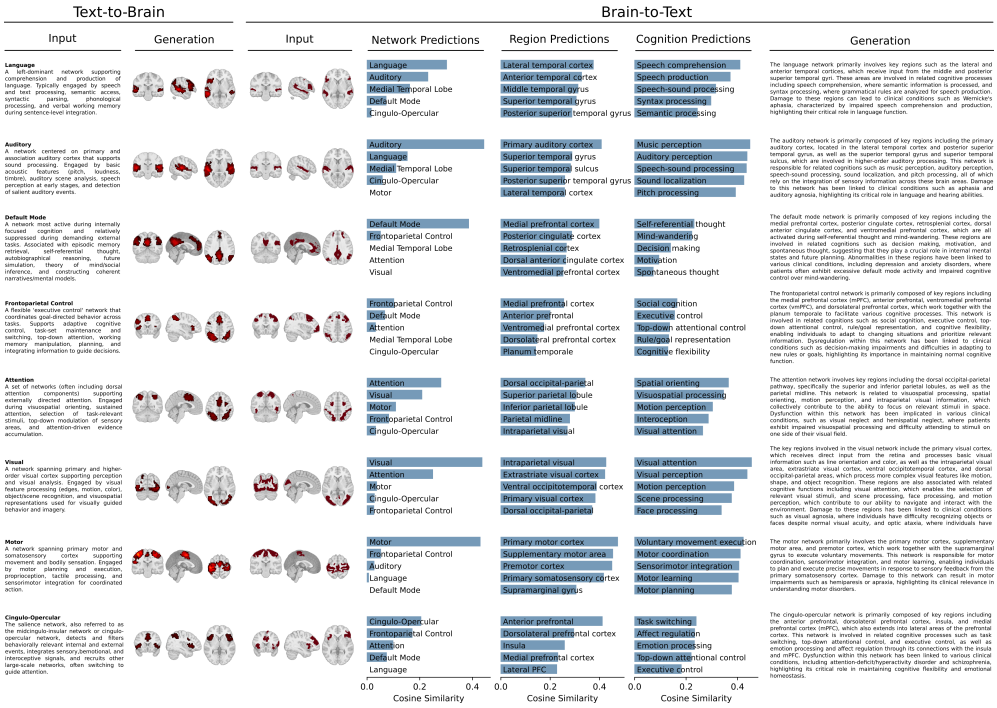


Fig. 2 Networks: text-to-brain and brain-to-text generation. The inputs to generation were labeled atlas networks and associated text. Text-to-brain generation used text description of networks to generate a set of activation maps. Input text description intentionally omits anatomical regions to increase the difficulty of the task. Brain-to-text generation used atlas images to generate interpretations of the brain map. An intermediate step in brain-to-text generation is predicting terms related to networks, regions, and cognition (blue).

The network dataset [1] was used to demonstrate text-to-brain and brain-to-text generative models, because it provides some degree of ground truth, given that the collection of networks were labeled by experts (Fig. 2 for examples, Fig. 4b for quantitative metrics). The text-to-brain model transformed networks' definitions to the latent space using the encoder. The latent vectors were then transformed to neuroimages using the decoder. The brain-to-text model generates text related to a neuroimage input from the network dataset. Concept labels are first predicted (Fig. 1b) and then passed to an LLM for text generation (Fig. S1). The LLM expands upon and provides additional context to concept predictions.

3.1.2 NeuroVault

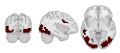
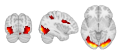
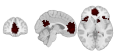
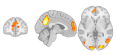
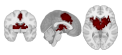
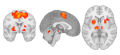
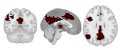
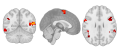
	Similarity		Text-to-Brain		Brain-to-Text	
Quality	Text	Images	True	Generated	True	Generated
High	0.86	0.74	Mapping visual symbols onto spoken language along the ventral visual stream			Visual Word Form Processing within the Dorsal Attention Network
High	0.86	0.75	Effects of propofol anesthesia on the processing of noxious stimuli in the spinal cord and the brain			Pain Perception: A Role for Insula Cortices in Modulating Intensity
Medium	0.47	0.49	Default mode contributions to automated information processing			Cognition through Medial Prefrontal Cortex Network Dynamics
Medium	0.51	0.45	Distinct regions of the striatum underlying effort, movement initiation and effort discounting			Motor Sequence Execution in the Sensorimotor Network
Low	0.12	0.04	Brain networks for confidence weighting and hierarchical inference during probabilistic learning			Attentional Control through Frontoparietal Network Dynamics
Low	0.08	0.16	Alpha-Band Phase Synchrony Is Related to Activity in the Fronto-Parietal Adaptive Control Network			Prefrontal Cortical Regions Contribute to Executive Control and Attentional Processes

Fig. 3 NeuroVault: text-to-brain and brain-to-text generation. Pairs of group-level activation maps and publication text (titles and abstracts) were taken from NeuroVault. Text was used to generate brain maps (right) and true brain maps were used to generate related text (right). Quality of generations are measured with cosine similarity between true and generated (left).

Generative neuroimage-to-text and text-to-neuroimage models were evaluated on the neurovault test set, which provides unthresholded brain maps not in the PubMed training set. 5-fold cross-validation was performed. Qualitative results are shown in Fig. 3, where each row in this figure corresponds to one group-level activation map and study from neurovault. Titles and abstracts from neurovault were used to generate 3D brain maps; only true and predicted (generated) titles are shown. Performance for both text-to-brain and brain-to-text models is quantified in Fig. 4c, where the area under the recall@k curve was between 0.78 and 0.80, indicating performance well above chance.

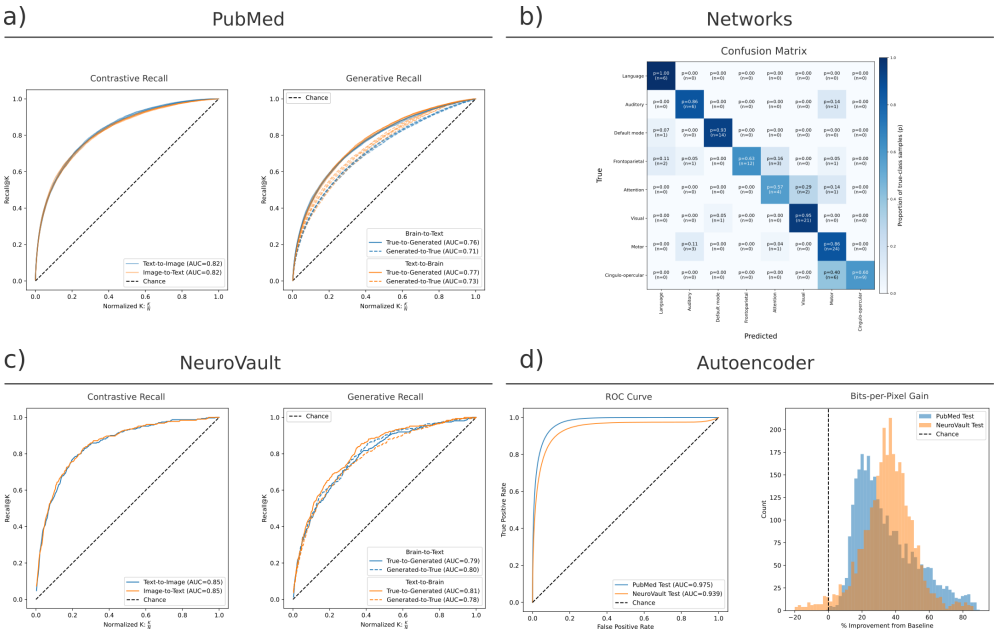


Fig. 4 Model Performance. Results are presented for the neuroimage autoencoder, contrastive models, and generative models. Recall@k and receiver operating characteristic curves (ROC) have expected random performance of area-under-the-curve (AUC=0.5), along the diagonal. Perfect performance is AUC=1.0. Recall@k measures how often the true pair is in the top-k most cosine similar pairs. **a) PubMed.** Contrastive (AUC=0.81) and generative (AUC=0.71) recall curves. Each trace is performance from 5-fold cross-validation. **b) Network classification.** Multi-class predictions, from `argmax` of cosine similarity between contrastive embeddings, as a confusion matrix. The diagonal is correct classifications (dark blue). Counts and proportion of true-sample class are given per cell. **c) NeuroVault.** Contrastive (AUC=0.85) and generative recall (AUC=0.78) curves. **d) Autoencoder.** Binary voxel performance (AUCROC=0.94) and bits-per-pixel gain from a naive baseline.

3.2 Contrastive Model

3.2.1 Network Labeling

Cosine similarity between text-image embeddings was used to label 300 network images [1]. The argmax of cosine similarity is taken as the predicted label. The confusion matrix shows a vast majority of images were labeled correctly using this approach (Fig. 4b). Network images were embedded to a common latent space and then compared against latent text embeddings corresponding to networks, regions, and cognition. Example network images and associated text are shown in Fig. 2.

The network dataset [1] also contains unlabeled independent components (IC) for the Human Connectome Project and UK Biobank. Independent component analysis (ICA) of resting state networks has historically required manual classification of networks versus noise components [27]. Semi-supervised approaches also require manual classifcation for a classifier to learn from [28, 29]. These methods classifiy components as either noise or target components. The ICA components from the network dataset

are labeled automatically using the contrastive model Fig. S3, with more detailed, e.g., the names of various networks. Although noise is not directly modeled, noise components will have low similarity with a resting-state text corpus.

3.2.2 NeuroVault

The NeuroVault [21] test set was used to test the contrastive model. This shows that the model generalizes from coordinate tables during training to group-level activation maps at inference. Recall@k measures how often the true neuroimage-text pair was found in the top-k most similar pairs (Fig. 4b, left) [16]. Similar to AUROC, the area under the curve is expected to be 0.5 for a randomly initialized, untrained model. After training with the PubMed set, the AUC reached 0.85, again showing performance well above chance.

3.3 Autoencoder

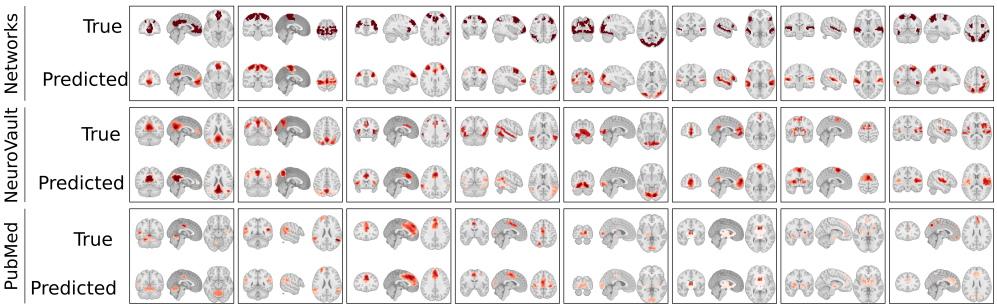


Fig. 5 Autoencoder test set reconstructions. Example reconstructions for the networks, NeuroVault, and PubMed datasets. The top in each box is the true image. The autoencoder encodes the images to latent space, then decodes the latent space to recreate the image. The reconstructed predictions are shown below each true image. All images were randomly selected from the test sets.

The first step in training is to learn an autoencoder that can reconstruct neuroimages. This was performed on the PubMed training set. Each MNI space activation map is encoded to a 368-dimensional latent vector then decoded back to MNI space. This latent vector space is the target for encoding text, resulting in text-to-brain decoding. The autoencoder was trained on binarized targets and predicts logits, e.g., each voxel was modeled as a binary classification problem. The model output is interpreted as the probability of each voxel being active. Performance is quantified by area under the receiver operating characteristic (AUROC = 0.975) and normalized negative log likelihood (bits/voxel) [30] that measures improvement relative to a naive baseline (Fig. 4d). The test set included network maps, NeuroVault images, and PubMed (Fig. 5).

4 Discussion

One of the most central dogmas of neuroscience is that of functional localization: the idea that specific cognitive and behavioral functions can be localized to specific brain regions. This view has been updated to a richer, network-level view of behavior and cognition. The history of human brain imaging is such that vast amounts of natural language data have been generated in attempts to describe the populations from whom brain images were recorded, under what experimental conditions, with expert narratives describing the resulting brain maps. Together, these texts and images provide a rich set of data upon which we collectively build novel hypotheses about the neural basis of behavior, cognition, and disease.

Here, we introduce NeuroVLM, a generative AI framework that can serve as an expert collaborator in this process of narrative building and hypothesis generation. This generative approach builds upon past works that focused on text-to-brain models, constrained to a finite, predefined corpus [3, 11]. The text-to-brain model was expanded to BERT-based text encoders [31]. More recent efforts have trained contrastive models to rank similarity between text-brain pairs using LLM embeddings [16, 17].

NeuroVLM is an integrative framework (Fig. 1a) that is capable of generating text from neuroimages or neuroimages from text. This framework relies on encoding to, and decoding from, a common latent space. This general framework is not dependent on the choice of encoders or decoders. Instead, we provide a neuroimage autoencoder, a model to project text embeddings to the shared latent space, and an interpretable method to generate text from neuroimages. These models were trained on 30,000 publications and extend to group-level statistical maps [21] and network atlases [1]. Use cases include generating atlases or maps given a text corpus, generating text interpretations of 3D brain maps, labeling networks with coactive brain regions, finding publications most related to a statistical map query, or finding statistical maps most related to a text query.

A central limitation to our approach is that our training neuroimages are not native statistical maps, but proxy maps reconstructed from reported peak coordinates and smoothed with a fixed kernel. Coordinate tables are not intended to be a sufficient statistic for the underlying effects — they omit subthreshold structure, cluster shape, effect-size information, variance, and many study-specific details such as contrast definitions, preprocessing choices, and sample characteristics. This choice is pragmatic for scale, but it bounds the quality of what the model can learn.

Another limitation involves topics that are under-powered and under-studied. While common networks and widely used terminology are well-represented, rare constructs, emerging topics, or underrepresented populations can be mapped less reliably. For more niche, specific topics, additional fine-tuning is recommended.

In addition, it is important to emphasize that generative models often can introduce confident predictions that go beyond the evidence. Using NeuroVLM, brain maps will be generated from any text input, no matter how ludicrous. For example, "show me the hot dog areas of the brain," will result in beautiful 3D brain images – largely regions related to gustation and reward, as it turns out (so perhaps not entirely ludicrous). Nevertheless, NeuroVLM should be viewed as an AI collaborator – a tool for organizing and navigating the literature, suggesting candidate networks, concepts, and related

papers – rather than as an oracle that provides mechanistic explanations or ground truth maps.

Despite these caveats, our work suggests that large-scale, paired text-neuroimage corpora are sufficient to learn a useful shared representation that supports practical tasks such as retrieval, labeling, and coarse generation. Our strongest limitations involve access to neuroimaging data rather than model architecture. Contributions to open-access repositories, such as NeuroVault [21] and OpenNeuro [32], will improve the future generations of multimodal neuroscience models; or more simply: higher quality, expertly labeled and annotated data can yield better results [33].

Looking ahead, we believe that it is critical to strengthen the ecosystem around standardized benchmarks and reproducible evaluation. Network labeling and recall-style retrieval provide accessible demonstrations, but they do not fully measure neuroscientific correctness, robustness to analysis pipelines, or calibration under dataset shift. Progress will benefit from community benchmarks that specify train-test splits, map types, preprocessing, and query sets, and that evaluate multiple axes of performance: retrieval utility, interpretability, out-of-distribution generalization, and failure modes. Open model checkpoints, transparent preprocessing, and reproducible training scripts should be prioritized.

NeuroVLM frames the neuroimaging literature as a bidirectional link between how studies are described in language and where effects are reported in the brain. This enables a single framework to support retrieval, labeling, and generation from tens-of-thousands of paired examples. These models can help organize a fragmented literature into a searchable, cross-modal interface: a way to move from maps to candidate concepts and related papers, and from text queries to spatial priors and exemplar results. By making these capabilities accessible with modular components and modest compute, this work reduces the barrier to using modern multimodal methods for everyday neuroscientific workflows.

Our results underscore that the limiting factor for biological fidelity is often the structure of what the field reports and shares. When the community primarily provides peak coordinates, models inherit the constraints of peak-coordinate reconstructions; in contrast, sharing more data, such as full statistical maps, contrast definitions, and rich metadata, creates a virtuous cycle and allows models to learn quantitative, uncertainty-aware representations that better reflect the evidence [33].

NeuroVLM is a step toward computational tools that augment, rather than replace, expert interpretation. Its most immediate value is as a literature navigation and hypothesis-generation aid that can surface relevant work, highlight commonalities across maps and descriptions, and provide interpretable suggestions. The broader opportunity is to couple these models with standardized benchmarks and richer open data so that future systems are reliable, reproducible, and grounded in the full statistical record of human neuroimaging research.

Supplementary information.

5 Generation

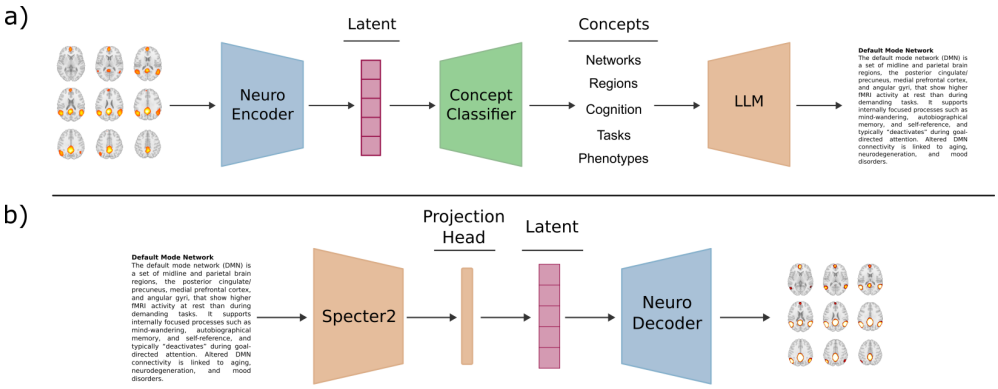


Fig. S1 Generating and interpreting neuroimages. a) Brain-to-text generation is used to predict text interpretations of input brain maps. An activation map is first encoded to latent vector (pink). Text labels are then predicted from the latent vector, either through a cosine similarity lookup or through a classification model (green). The top most similar concepts are then passed to an LLM (orange) to generated a summary of. The concept classifier provides a human-interpretable intermediate layer. Alternatively, the latent vector (pink) may be directly mapped to the token embedding space of the LLM through supervised learning. **b)** A text query is first encoded with Specter2, a fixed, pre-trained text encoder. The projection head is a small, trainable extension of the text encoder that maps from Specter2's embedding space to the bi-modal, text-neuoimage embedding space (pink). This latent vector is decoded (blue) to generated activation maps.

6 Preprocessing

6.1 Coordinate Tables

Coordinate tables (n=30826) were collected from PubMed and NeuroSynth Compose databases. Coordinates were convolved with a 9mm Activation Likelihood Estimation (ALE) kernel. The kernel was normalized to values between zero and one. The zero-to-one range allows using binary cross-entopy loss when training the autoencoder. The MNI-space convolutional product, $\mathbf{X}$, was clamped to values between zero and one. The 512-dimensional DiFuMo atlas, $\mathbf{P}$, were used to transform the convolved ALE products (Fig. S2). This is in contrast to using the 512-dimensional embedding directly, as in [16, 17]. Instead, the autoencoder is trained to learn a data-driven, low-dimensional embedding space.

$$\hat{\mathbf{X}} = \mathbf{X}\mathbf{P}\mathrm{diag}(s)^{-1}\mathbf{P}^{\top}, \text{where } s_j = \sum_v \mathbf{P}_{vj}$$

The DiFuMo reconstructions were resampled to 4mm isotropic voxel space, matching [11]. Masking and flattening produced a 28,542 dimensional vector per image.

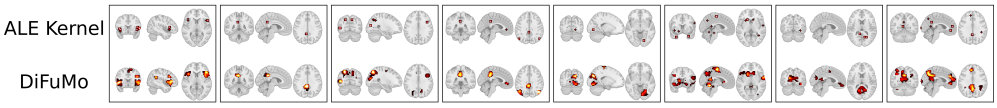


Fig. S2 ALE products transformed to DiFuMo. Points, from coordinate tables were first convolved with an activation likelihood estimation (ALE) kernel (top). The image output from ALE kernel convolution were reconstructed with DiFuMo (bottom).

6.2 Text

Text included titles and abstracts (n=30826) from PubMed publications. All text is paired to a DOI. Text was embedded using Specter2 [25]. This produces a 768-dimensional vector per publication. Specter2 was trained with contrastive loss, promoting ranking similarity between pairs. These embeddings almost collapse to a mean. This results in high cosine similarity for unrelated pairs, e.g. the expected cosine similarity distribution is shifted upwards by a constant. For text-to-brain decoding, the embeddings are demeaned with an empty text embedding, shifting the expected value of cosine similarity between random pairs to zero. Additionally, text embeddings were unit normalized, as contrastive loss is scale insensitive.

6.3 NeuroVault

NeuroVault images ($n = 4553$) span several group-level map types reflecting different analysis outputs, including inferential statistics maps: t-, z-, and F-maps (voxelwise test statistics for a contrast model) and p-maps (the corresponding voxelwise significance values). We also included beta maps (univariate GLM estimates and multivariate/decoding weight maps) as well as ROI/masks (binary/probabilistic inclusion maps) and parcellations (discrete region labels), which define spatial subsets or atlas assignments rather than continuous statistics.

Activation maps were cluster-thresholded. For each map, we applied a high-intensity threshold at the 99th percentile and removed small clusters using a minimum cluster extent of 50 voxels. The surviving suprathreshold clusters were then re-vectorized with the same masker into a fixed-length feature vector, $d = 28542$, and used in testing. Finally, we binarized the resulting vectors (presence/absence of surviving voxels) because the absolute scaling of values across map types (e.g., t-, z-, p-, F-, and beta maps) is not directly comparable; binarization yields a uniform support that emphasizes spatial localization over intensity.

NeuroVault text was limited to titles and abstracts, since per-image contrast definitions were often noisy or missing. Since multiple images were paired to each title-abstract pair, the contrastive model was used to select the most similar image. This inflates scores and provides an upper performance bound. All $n = 4553$ images

Table 1 Autoencoder architecture (layer-wise).

Layer	Module	Input → Output	Activation
Encoder 1	Linear	28542 → 1024	ReLU
Encoder 2	Linear	1024 → 512	ReLU
Encoder 3	Linear	512 → 384	None
Decoder 1	Linear	384 → 512	ReLU
Decoder 2	Linear	512 → 1024	ReLU
Decoder 3	Linear	1024 → 28542	None

Shapes omit batch dimension.

were used when evaluating the autoencoder, since the autoencoder does not require text-image pairs.

7 Autoencoder

7.1 Architecture

After resampling to 4mm^3 voxel space, masking, flattening and each neuroimage is a 28,542 dimensional vector. The exact layers sizes and activation functions are provided in Tab. 1. The encoder maps from 28,542 to 384, then the decoder maps from 384 back to 28,542. ReLU activations are used. The model can be thought of as a single masked convolution per layer, with a kernel size equal to to input size. This simplifies handling sparsity from outside-of-brain voxels. Alternatively, convolutional and transposed convolutional layers with a smaller kernel could be explored in future work.

7.2 Binary Cross-Entropy Loss

The autoencoder was trained with BCE-with-logits loss. Let $\mathbf{x}^{(i)} \in [0,1]^D$ denote a masked and flattened activation map with $d = 28542$. The decoder produces logits $\mathbf{z}^{(i)} \in \mathbb{R}^d$, that are transformed to probabilities using the sigmoid function, $\sigma(u)$.

$$\hat{\mathbf{x}}^{(i)} = \sigma\Big(\mathbf{z}^{(i)}\Big), \qquad \sigma(u) = \frac{1}{1+\exp(-u)} \tag{1}$$

We minimize the binary cross-entropy between $\mathbf{x}^{(i)}$ and $\hat{\mathbf{x}}^{(i)}$ using the numerically stable BCE-with-logits form:

$$\mathcal{L}_{\text{BCELogits}}\Big(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}\Big) = -\frac{1}{D}\sum_{j=1}^{D}\Big[x_j^{(i)} \log \sigma(z_j^{(i)}) + \big(1 - x_j^{(i)}\big) \log\big(1 - \sigma(z_j^{(i)})\big)\Big] \tag{2}$$

$$= \frac{1}{D}\sum_{j=1}^{D}\Big(\max(z_j^{(i)}, 0) - z_j^{(i)} x_j^{(i)} + \log\big(1 + \exp(-|z_j^{(i)}|)\big)\Big). \tag{3}$$

Table 2 Projection head architecture (layer-wise).

Layer	Module	Input → Output	Activation
Text 1	Linear	768 → 512	ReLU
Text 2	Linear	512 → 384	None
Neuro 1	Linear	384 → 384	ReLU
Neuro 2	Linear	384 → 384	None

Shapes omit batch dimension.

The overall reconstruction loss is the minibatch average over B examples:

$$\mathcal{L}_{\text{recon}} = \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_{\text{BCELogits}}\left(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}\right). \tag{4}$$

7.3 Training

The autoencoder was trained for 100 epochs using AdamW with a learning rate of 5×10^{-5} and a batch size of 256. Reconstruction was optimized with binary cross-entropy loss with logits as described above.

8 Projection Head

8.1 Architecture

There are three types of projection heads. The first two are used for transform 768 dimensional Specter2 embeddings to 384 dimensions. These two projection heads were the same size but trained on different targets. One was trained to map latent text to the autoencoders latent space, providing text-to-brain decoding. The other text projection head was learned alongside the neuro projection head, optimized with the contrastive objective, InfoNCE loss. The neuro projection head maps from the autoencoder latent space to the bi-modal latent space. Tab. 2 contains the exact layers sizes.

8.2 InfoNCE Loss

Contrastive projection heads were trained with Info Noise Contrastive Estimation (InfoNCE) loss. Let $\mathbf{I} \in \mathbb{R}^{B \times d}$ and $\mathbf{T} \in \mathbb{R}^{B \times d}$ represent a batch of image and text embeddings, respectively, where B is the batch size and d is the embedding dimension. We used $B = 2048$ and ℓ_2-normalize each embedding for $i \in \{1, \ldots, B\}$:

$$\tilde{\mathbf{i}}_i = \frac{\mathbf{i}_i}{\|\mathbf{i}_i\|_2}, \qquad \tilde{\mathbf{t}}_i = \frac{\mathbf{t}_i}{\|\mathbf{t}_i\|_2}, \tag{5}$$

We then compute the pairwise cosine similarity matrix:

$$\mathbf{S} \;=\; \frac{1}{\tau} \tilde{\mathbf{I}} \tilde{\mathbf{T}}^{\top} \in \mathbb{R}^{B \times B}, \tag{6}$$

The temperature parameter was set to: $\tau = 0.07$. The correct text-image pair for each index i is the diagonal entry S_{ii}, all off-diagonal entries are mismatched pairs, i.e. text-image pairs from mismatched studies. Thus, target labels are the identity matrix. We optimize a symmetrized cross-entropy loss over image-to-text ($\mathcal{L}_{\text{i2t}}$) and text-to-image ($\mathcal{L}_{\text{t2i}}$) retrieval. The final loss is the mean of the two.

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\mathbf{S}_{ii})}{\sum_{k=1}^{B}\exp(\mathbf{S}_{ik})} \tag{7}$$

$$\mathcal{L}_{\text{t2i}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\mathbf{S}_{ii})}{\sum_{k=1}^{B}\exp(\mathbf{S}_{ki})} \tag{8}$$

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{2}\left(\mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}\right). \tag{9}$$

8.3 Training

The projection head for text-to-brain generation was trained for 200 epochs with a batch size of 1024 using AdamW with a learning rate of 5×10^{-5}. Optimization of this text-to-brain model used a BCE-with-logits objective. For joint text–image contrastive alignment, we trained paired projection heads using a symmetric InfoNCE contrastive objective with temperature, $\tau = 0.07$. Models were optimized with AdamW for 300 epochs using a batch size of 2048 and learning rate of 1×10^{-5}.

8.4 Independent Component Labeling

Fig. S3 Independent Component Labeling. Automated labeling of resting-state components from the Human Connectome Project and UK Biobank. This has historically required manual labels.

References

[1] Kong, R. *et al.* A network correspondence toolbox for quantitative evaluation of novel neuroimaging results. *Nature Communications* **16**, 2930 (2025). URL https://www.nature.com/articles/s41467-025-58176-9.

[2] Yarkoni, T., Poldrack, R. A., Van Essen, D. C. & Wager, T. D. Cognitive neuroscience 2.0: building a cumulative science of human brain function. *Trends in*

Cognitive Sciences **14**, 489–496 (2010). URL http://dx.doi.org/10.1016/j.tics.2010.08.004.

[3] Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods* **8**, 665–670 (2011). URL https://www.nature.com/articles/nmeth.1635.

[4] Turner, B. O., Paul, E. J., Miller, M. B. & Barbey, A. K. Small sample sizes reduce the replicability of task-based fmri studies. *Communications Biology* **1** (2018). URL http://dx.doi.org/10.1038/s42003-018-0073-z.

[5] Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020). URL http://dx.doi.org/10.1038/s41586-020-2314-9.

[6] Markello, R. D. *et al.* neuromaps: structural and functional interpretation of brain maps. *Nature Methods* **19**, 1472–1479 (2022). URL http://dx.doi.org/10.1038/s41592-022-01625-w.

[7] Voytek, B. The data science future of neuroscience theory. *Nature Methods* **19**, 1349–1350 (2022). URL http://dx.doi.org/10.1038/s41592-022-01630-z.

[8] Voytek, J. B. & Voytek, B. Automated cognome construction and semi-automated hypothesis generation. *Journal of Neuroscience Methods* **208**, 92–100 (2012). URL https://linkinghub.elsevier.com/retrieve/pii/S0165027012001513.

[9] Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019). URL https://www.nature.com/articles/s41586-019-1335-8.

[10] Hansen, J. Y. & Misic, B. Integrating and interpreting brain maps. *Trends in Neurosciences* **48**, 594–607 (2025). URL https://linkinghub.elsevier.com/retrieve/pii/S0166223625001249.

[11] Dockès, J. *et al.* NeuroQuery, comprehensive meta-analysis of human brain mapping. *eLife* **9**, e53385 (2020). URL https://elifesciences.org/articles/53385.

[12] Vaswani, A. *et al.* *Attention is all you need*, NIPS'17, 6000–6010 (Curran Associates Inc., Red Hook, NY, USA, 2017).

[13] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018). URL https://arxiv.org/abs/1810.04805. Version Number: 2.

[14] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. *et al.* Improving language understanding by generative pre-training (2018).

[15] Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision (2021). URL https://arxiv.org/abs/2103.00020. Version Number: 1.

[16] Meudec, R., Ghayem, F., Dockès, J., Wassermann, D. & Thirion, B. in *NeuroConText: Contrastive Text-to-Brain Mapping for Neuroscientific Literature* (eds Linguraru, M. G. *et al.*) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Vol. 15003 325–335 (Springer Nature Switzerland, Cham, 2024). URL https://link.springer.com/10.1007/978-3-031-72384-1_31. Series Title: Lecture Notes in Computer Science.

[17] Peraza, J. A. *et al.* NiCLIP: Neuroimaging contrastive language-image pretraining model for predicting text from brain activation images (2025). URL http://biorxiv.org/lookup/doi/10.1101/2025.06.14.659706.

[18] Cohan, A., Feldman, S., Beltagy, I., Downey, D. & Weld, D. S. SPECTER: Document-level Representation Learning using Citation-informed Transformers (2020). URL https://arxiv.org/abs/2004.07180. Version Number: 4.

[19] Kent, J. *et al.* 141. Neurosynth Compose: A Free an Open Platform for Precise Large-Scale Neuroimaging Meta-Analysis. *Biological Psychiatry* **95**, S156–S157 (2024). URL https://doi.org/10.1016/j.biopsych.2024.02.376. Publisher: Elsevier.

[20] Poldrack, R. A. *et al.* The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Frontiers in neuroinformatics* **5**, 17 (2011).

[21] Gorgolewski, K. J. *et al.* NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics* **9** (2015). URL http://journal.frontiersin.org/article/10.3389/fninf.2015.00008/abstract.

[22] Dadi, K. *et al.* Fine-grain atlases of functional modes for fMRI analysis. *NeuroImage* **221**, 117126 (2020). URL https://linkinghub.elsevier.com/retrieve/pii/S1053811920306121.

[23] Beltagy, I., Lo, K. & Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv preprint* (2019). URL https://arxiv.org/abs/1903.10676. Publisher: arXiv Version Number: 3.

[24] Li, J., Li, D., Savarese, S. & Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023). URL https://arxiv.org/abs/2301.12597.

[25] Singh, A., D'Arcy, M., Cohan, A., Downey, D. & Feldman, S. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations (2023). URL http://arxiv.org/abs/2211.13308. ArXiv:2211.13308 [cs].

[26] Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models (2021). URL https://arxiv.org/abs/2106.09685.

[27] Griffanti, L. *et al.* Hand classification of fmri ica noise components. *NeuroImage* **154**, 188–205 (2017). URL http://dx.doi.org/10.1016/j.neuroimage.2016.12.036.

[28] Salimi-Khorshidi, G. *et al.* Automatic denoising of functional mri data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage* **90**, 449–468 (2014). URL http://dx.doi.org/10.1016/j.neuroimage.2013.11.046.

[29] Pruim, R. H. *et al.* Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data. *NeuroImage* **112**, 267–277 (2015). URL http://dx.doi.org/10.1016/j.neuroimage.2015.02.064.

[30] Oord, A. v. d. *et al.* Conditional Image Generation with PixelCNN Decoders (2016). URL https://arxiv.org/abs/1606.05328. Version Number: 2.

[31] Ngo, G. H., Nguyen, M., Chen, N. F. & Sabuncu, M. R. in *Text2Brain: Synthesis of Brain Activation Maps from Free-Form Text Query* (eds De Bruijne, M. *et al.*) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Vol. 12907 605–614 (Springer International Publishing, Cham, 2021). URL https://link.springer.com/10.1007/978-3-030-87234-2_57. Series Title: Lecture Notes in Computer Science.

[32] Markiewicz, C. J. *et al.* The OpenNeuro resource for sharing of neuroscience data. *eLife* **10**, e71774 (2021). URL https://elifesciences.org/articles/71774.

[33] Voytek, B. The virtuous cycle of a data ecosystem. *PLoS Comput Biol* **12**, e1005037 (2026). URL https://doi.org/10.1371/journal.pcbi.1005037.