General Artificial Intelligence (1)

SAIR-2-06: Foundation Model for Decoding Brain fMRI

Momiao Xiong

Society of Artificial Intelligence Research

BRAINLM:

A FOUNDATION MODEL FOR BRAIN ACTIVITY RECORDINGS

- Brain LanguageModel (BrainLM), a foundation model for brain activity dynamic trained on 6,700 hours of fMRI recordings. Utilizing self-supervised masked-prediction training
- BrainLM demonstrates proficiency in both fine-tuning and zero-shot inference tasks
- Fine-tuning allows for the prediction of clinical variables and future brain states.
- In zero-shot inference, the model identifies functional networks and generates interpretable latent representations of neural activity.
- a novel prompting technique, allowing BrainLM to function as an in silico simulator of brain activity responses to perturbations

- **Foundation models represent a new paradigm in artificial intelligence, shifting from narrow, task-specific training to more general and adaptable models**
- **the foundation model approach trains versatile models on broad data at scale, enabling a wide range of downstream capabilities via transfer learning.**

The foundation model approach trains versatile models on broad data at scale, enabling a wide range of downstream capabilities via transfer learning.

After pretraining, BrainLM supports diverse downstream applications via fine-tuning and zero-shot inference. We demonstrate BrainLM's capabilities on key tasks including prediction of future brain states, decoding cognitive variables, discovery of functional networks, and in silico perturbation analysis.

Recent Approach

Seeing Beyond the Brain: Conditional Diffusion Model with Sparse MaskedModeling for Vision Decoding

Zijiao Chen1* Jiaxin Qing2* Tiange Xiang3 Wan Lin Yue1 Juan Helen Zhou1†
1National University of Singapore, 2The Chinese University of Hong Kong, 3Stanford University

https://mind-vis.github.io (code)

Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity

Zijiao Chen, National University of Singapore, zijiao.chen@u.nus.edu
Jiaxin Qing, The Chinese University of Hong Kong, jqing@ie.cuhk.edu.hk
Juan Helen Zhouy, National University of Singapore, helen.zhou@nus.edu.sg

https://mind-video.com (Code)

Begin: Most materials come from the above first paper.

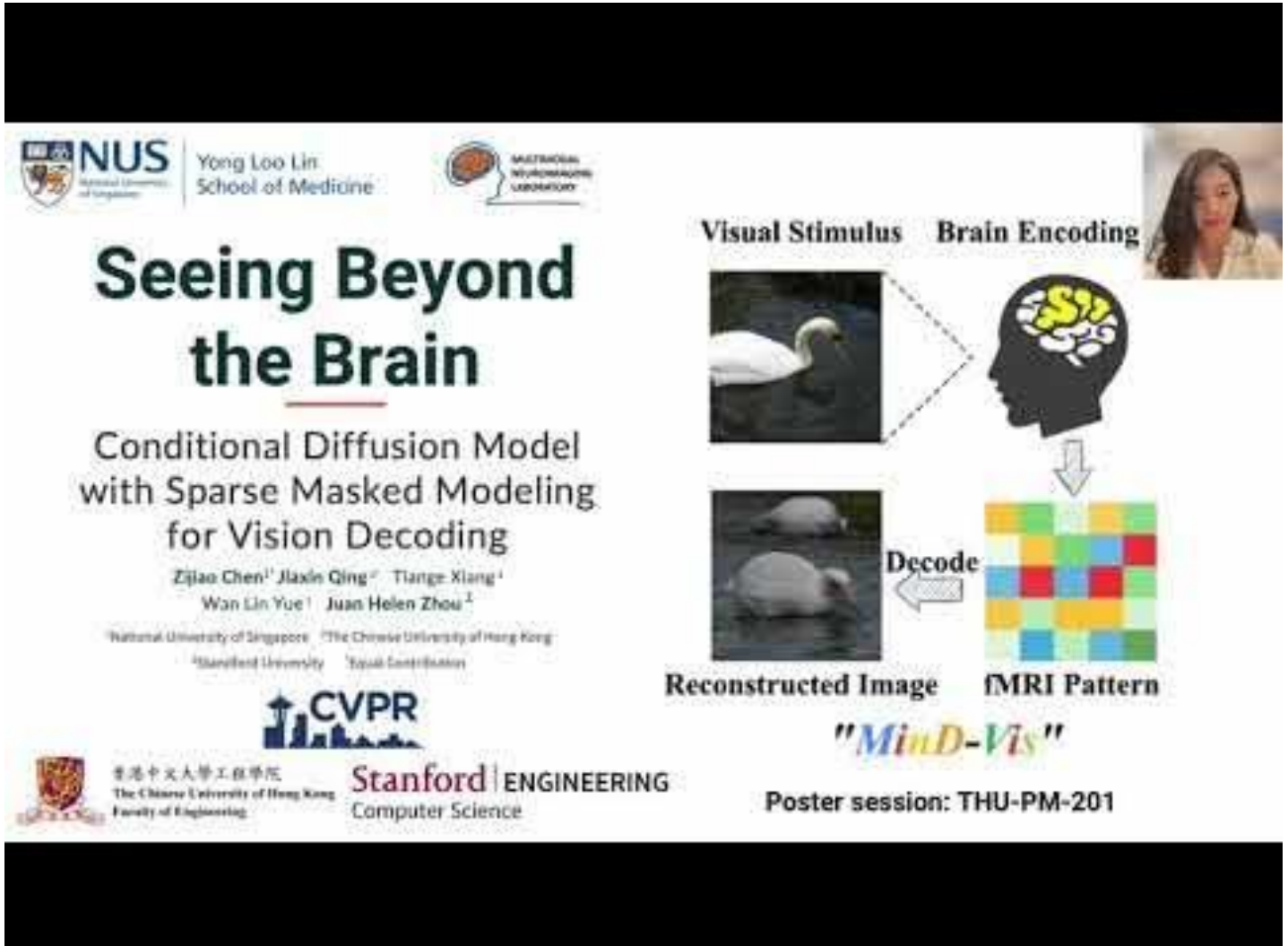
- Decoding visual stimuli from brain recordings aims to deepen our understanding of the human visual system and build a solid foundation for bridging human and computer vision through the Brain-Computer Interface (Next Time)

- MinD-Vis: Sparse Masked Brain Modeling with Double-Conditioned Latent Diffusion Model for Human Vision Decoding.

- Firstly, we learn an effective self-supervised representation of fMRI data using mask modeling

- Then by augmenting a latent diffusion model with double-conditioning for Human Vision Decoding

- **Human perception and prior knowledge are deeply intertwined in one's mind**
- **Our perception of the world is determined not only by objective stimuli properties but also by our experiences, forming complex brain activities underlying our perception.**



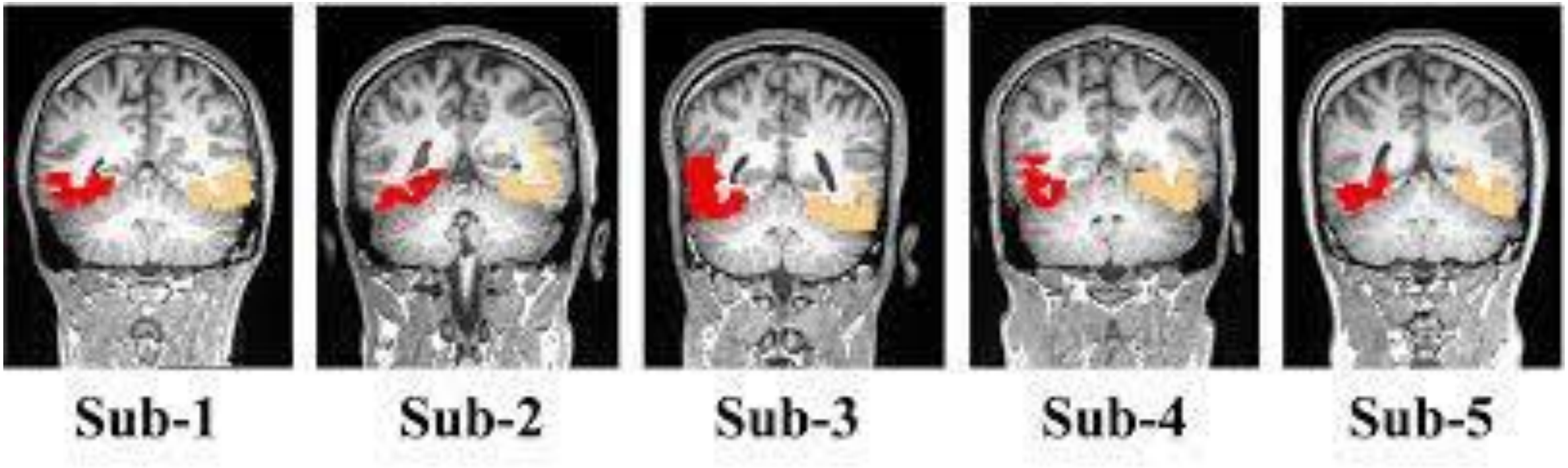


Figure 2. Individual Differences in Regions Responding to Visual Stimuli. Masks of the regions of interest activating during the same visual task differ in location and size across subjects. The primary visual cortex at the left (red) and the right (orange) hemisphere are shown.

aim to learn representations from a large-scale dataset with rich demographic compositions and relax the direct generation fromfMRI to conditional synthesis allowing for sampling variance under the same semantic category

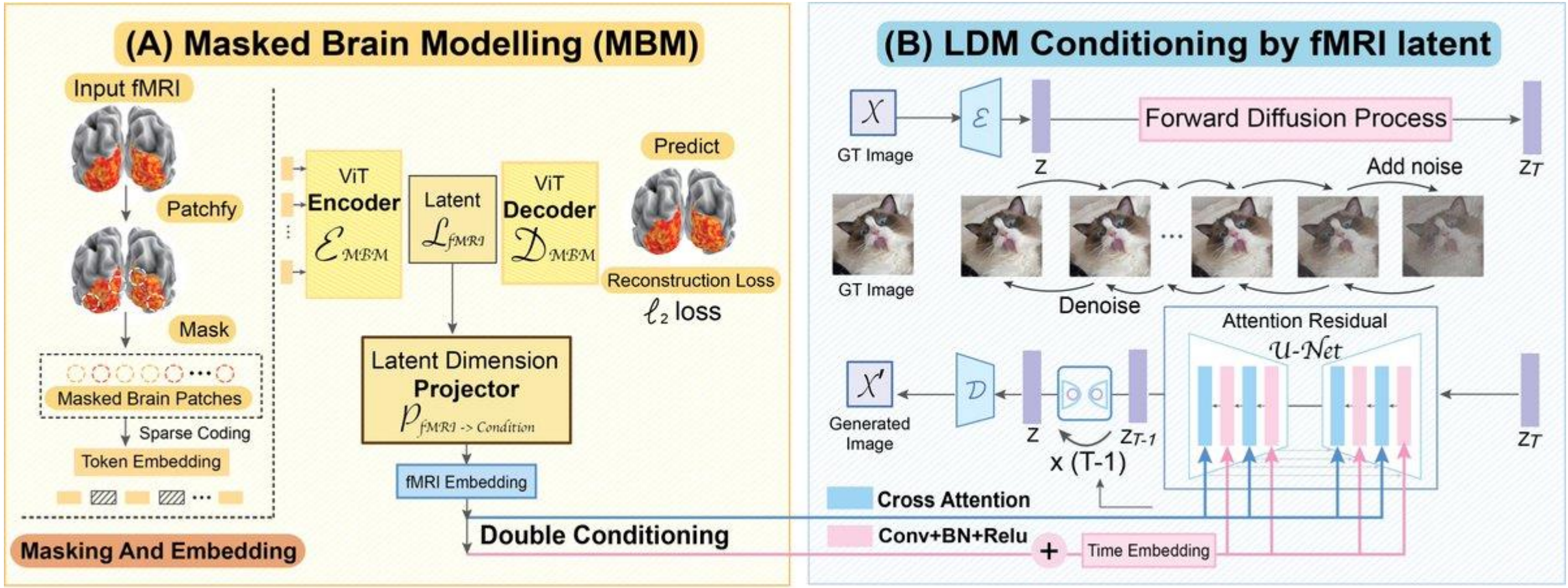
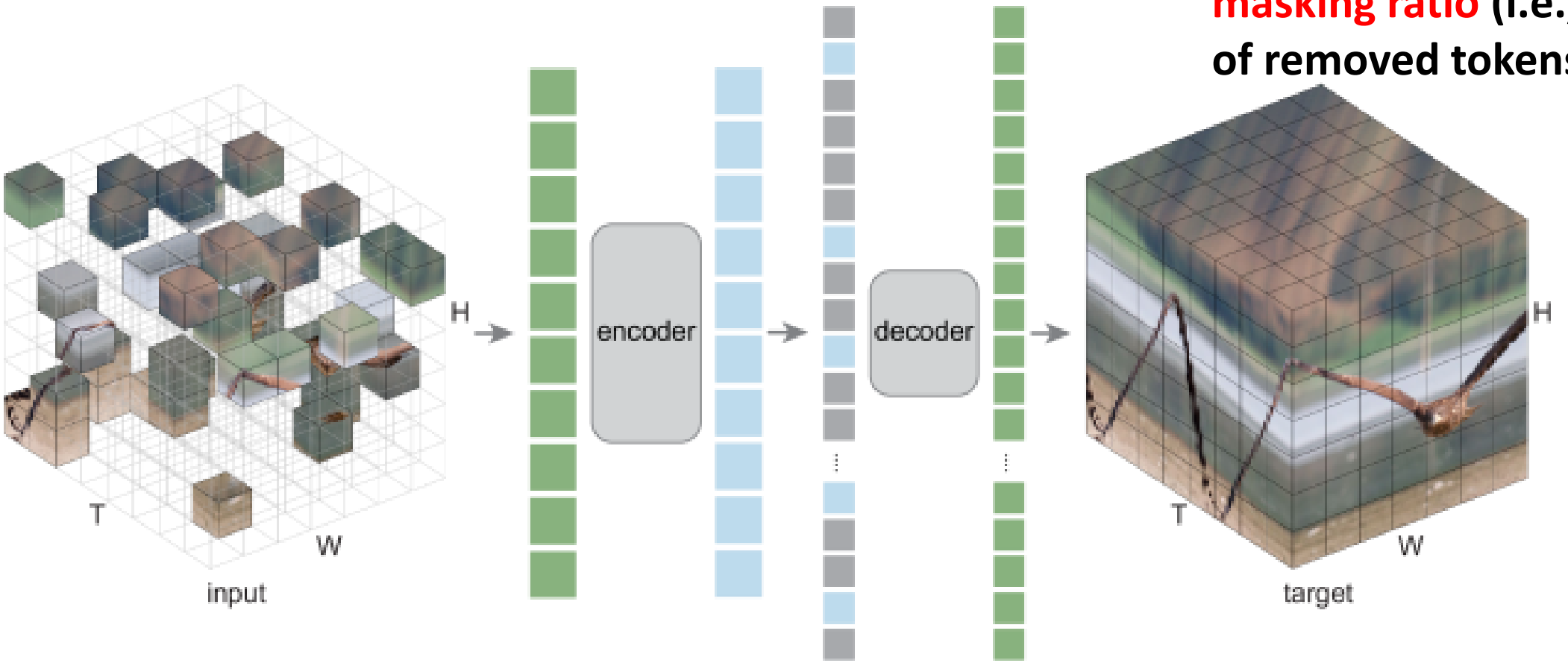


Figure 3. MinD-Vis. Stage A(left): Pre-train on fMRI with SC-MBM. We patchify, randomlymask the fMRI, and then tokenize them to large embeddings. We train an autoencoder (EMBM anDMBM) to recover the masked patches. Stage B (right): Integration with the LDMthrough double conditioning. We project the fMRI latent (LfMRI) through two paths to the LDMconditioning space with a latent dimension projector (PfMRI→Cond). One path connects directly to cross-attention heads in the LDM. Another path adds the fMRI latent to time embeddings. The LDMoperates on a low-dimensional, compressed version of the original image (i.e. image latent), however, the original image is used in this figure for illustrations.

predicts pixel values

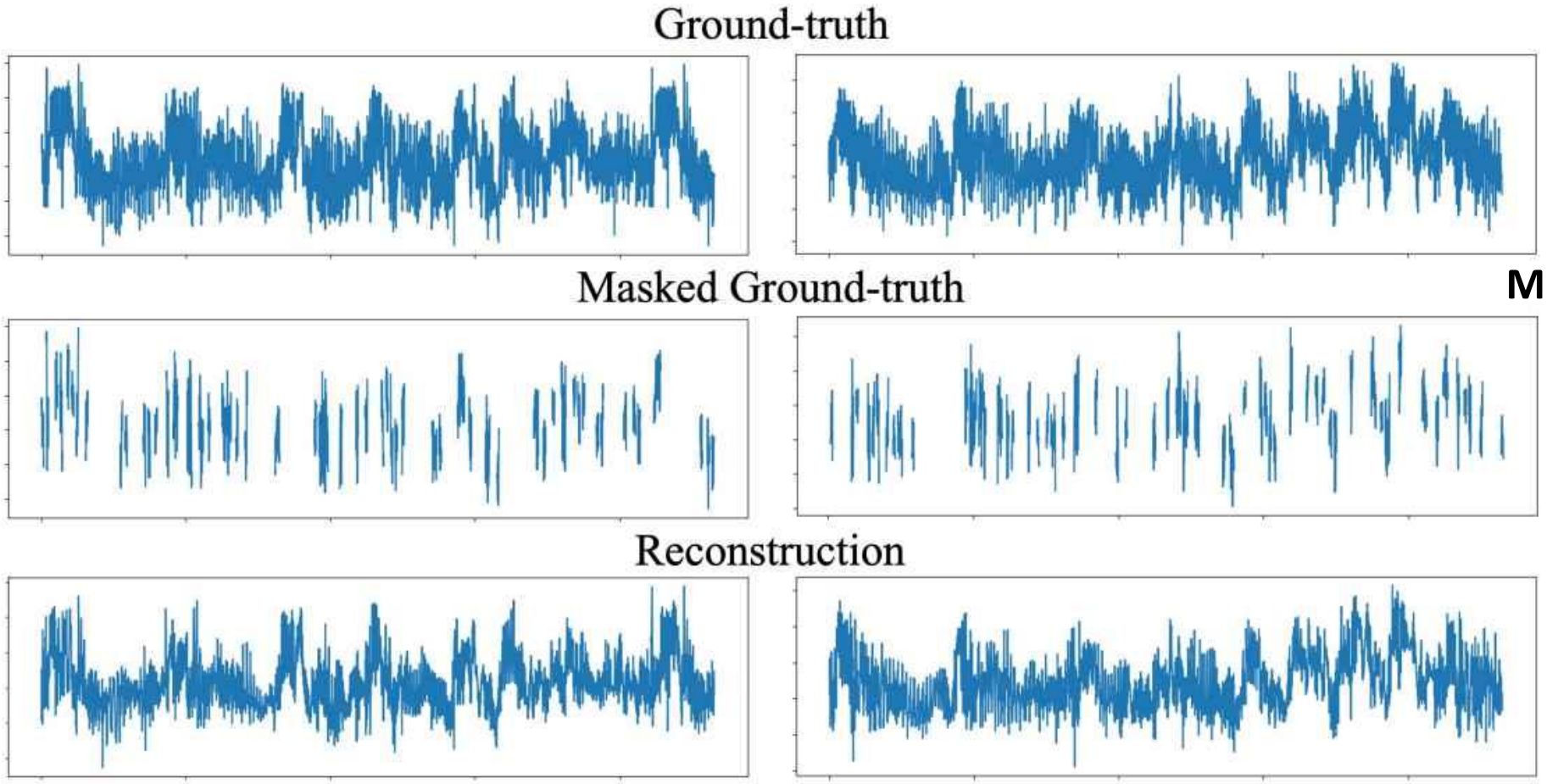
masking ratio (i.e., percentage of removed tokens)



The only spacetime-specific inductive bias is on embedding the patches and their positions; all other components are agnostic to the spacetime nature of the problem.

In particular, our encoder and decoder are both vanilla Vision Transformers [18] with no factorization or hierarchy, and our random mask sampling is agnostic to the spacetime structures

Stage A: Sparse-Coded MBM (SC-MBM)



Mask ratio 0.75; 4500 voxels

fMRI data can still be recovered even if a large portion is masked

Experiments

- **Datasets**

Three public datasets were used in this study:
Human Connectome Project (HCP) 1200 Subject Release [55];
Generic Object Decoding Dataset (GOD) [21]; and Brain,
Object, Landscape Dataset (BOLD5000) [5].

Our upstream pre-training dataset comprised fMRI data from HCP and GOD.

Combining these two, we obtained 136,000 fMRI segments
from **340 hours of fMRI scan**, which is, by far, the largest
fMRI pre-training dataset in the fMRI-image decoding task.

While the GOD is an fMRI-image paired dataset designed for fMRI-based decoding. The pairs in GOD were used for finetuning in our main analysis. The GOD consists of 1250 different images from 200 distinct classes, in which 1200 images were used as the training set, and the remaining 50 images were used as the testing set. The training set and testing set have no overlapping classes.

BOLD5000 dataset was used as the validation dataset in our study. It consists of 5254 fMRI-image pairs from 4916 distinct images, 113 images of which are used for testing. This is the first time that the BOLD5000 is used for fMRI decoding tasks.

Implementation

I pre-training model is similar to ViT-Large [10] with a 1D patch embedder.

a patch size of 16, embedding dimension of 1024, encoder depth of 24, and mask
.75 as our Full model setting with an ImageNet class-conditioned pre-trained LDM

on Metric

assification Accuracy

le trials, top-1 and top-5 classification accuracies were calculated in n−1 randomly selected
us the correct one

nception distance (FID)

9] is a commonly used metric to assess image generation quality. In our experiments, we
the FID between ground-truth images and generated images in the testing set.

Results

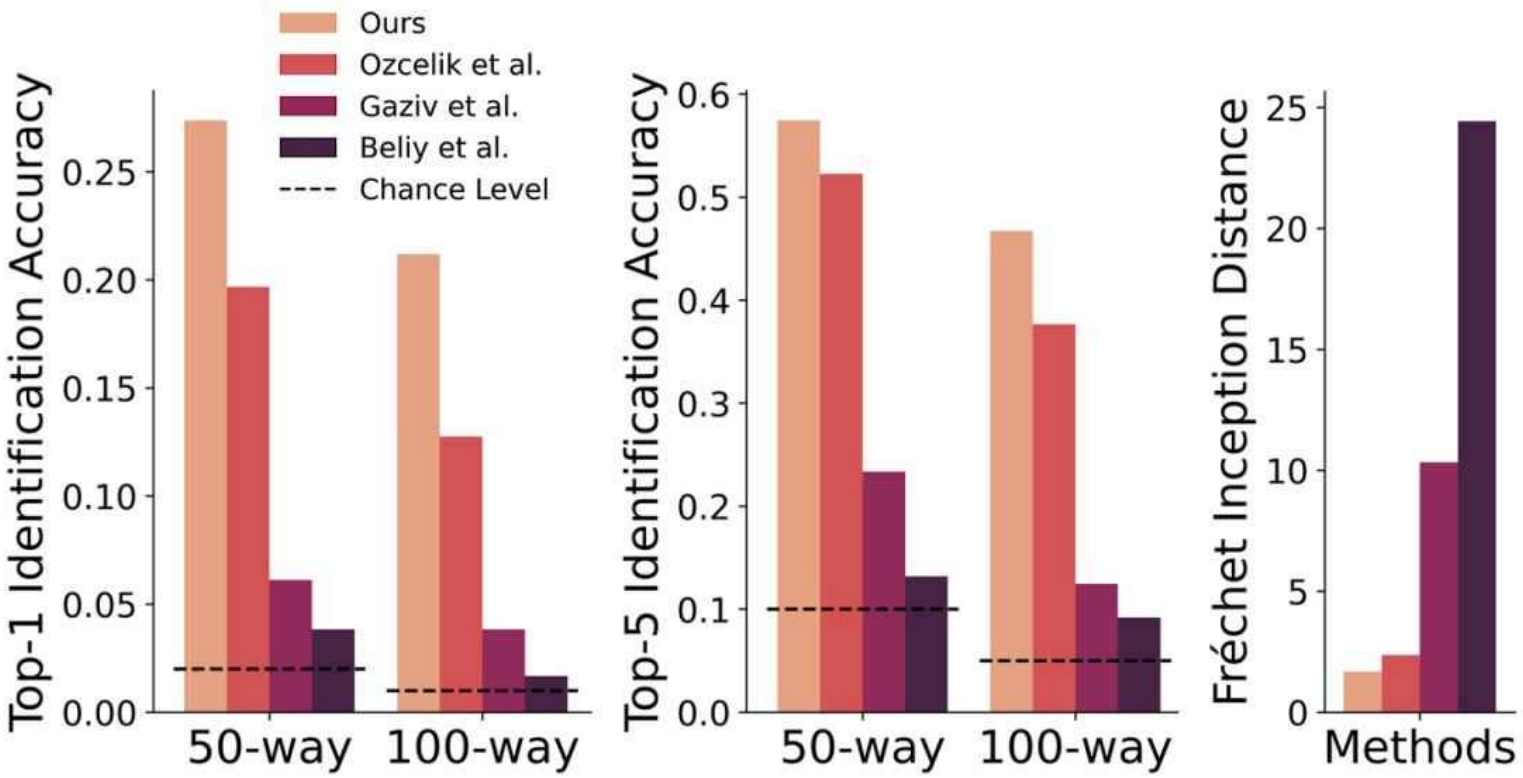


Figure 6. **Quantitative Performance Comparisons on GOD Test Set.** Performance is evaluated in terms of semantic correctness (*1000-trial n-way top-k classification accuracy; the higher the better*) and generation quality (*FID; the lower the better*).

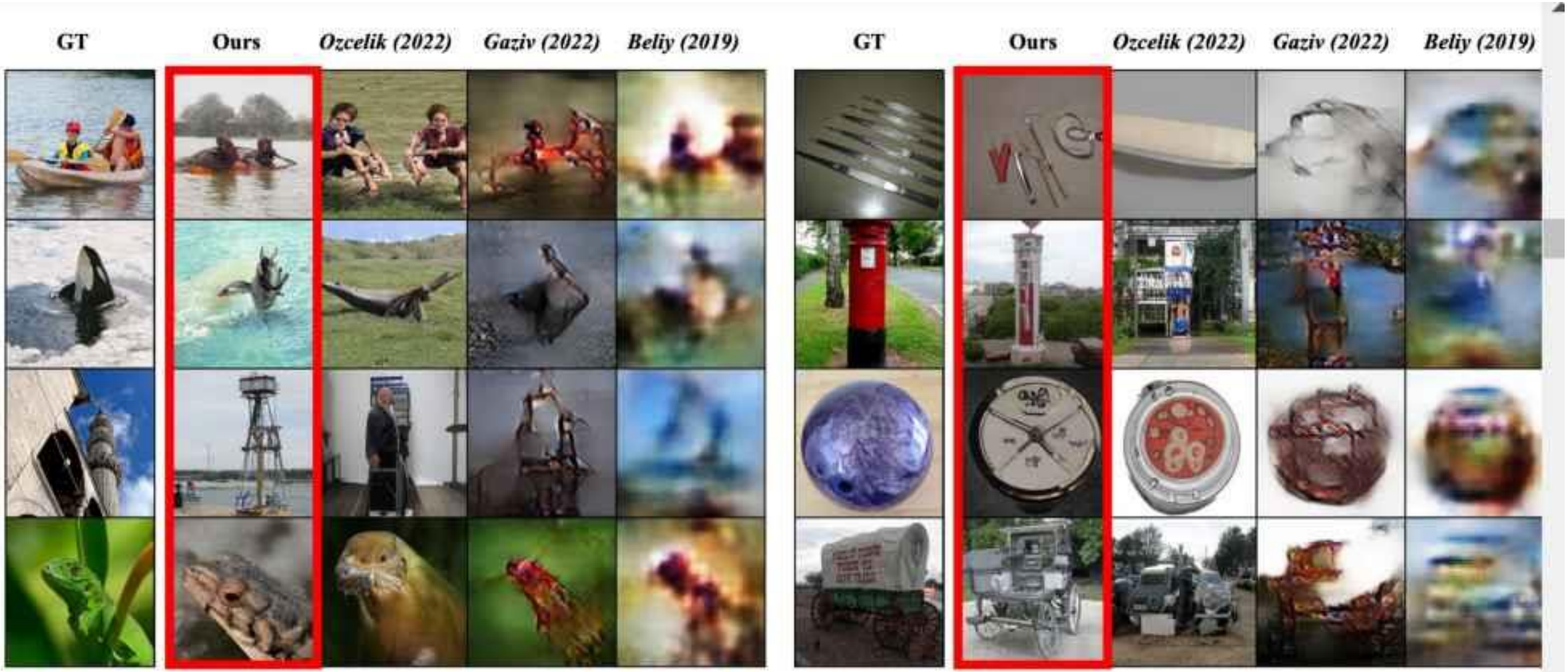


Figure 5. **Decoding Performance Comparisons on GOD Test Set.** The ground truth, images reconstructed by MinD-Vis and images reconstructe from three other methods are shown for comparison. MinD-Vis decoded the most accurate and plausible images with semantically similar details

our method generated plausible details such as water and waves in the first and second images

The image quality is also reflected by the FID, wherewe achieved **1.67** with our best samples, while Ozcelik et al. and others achieved **2.36**

Figure 7. **Generation Consistency of MinD-Vis**. Images generate by our method were consistent across different samplings trial sharing similar low-level features and semantics.

Model	Embedding Dim	Mask Ratio	Params	Acc (%)
Full	**1024**	**0.75**	**303M**	**23.9** ±3.00
1	w/o SC-MBM + same Encoder		**303M**	2.6 ±1.39
2	w/o SC-MBM + smaller Encoder		**25M**	3.4 ±0.86
3	**32**	0.75	0.3M	5.4 ±1.50
4	**64**	0.75	1.2M	6.9 ±1.10
5	**128**	0.75	4.7M	14.8 ±1.78
6	**256**	0.75	18.9M	15.9 ±1.70
7	**512**	0.75	75.6M	17.9 ±2.58
8	**768**	0.75	170M	17.7 ±1.42
9	**1280**	0.75	472M	15.5 ±3.83
10	1024	**0.35**	303M	19.6 ±3.40
11	1024	**0.45**	303M	20.0 ±1.89
12	1024	**0.55**	303M	18.1 ±2.87
13	1024	**0.65**	303M	21.7 ±3.61
14	1024	**0.85**	303M	16.1 ±1.00

† $p < 0.0001$ (purple); $p < 0.01$ (pink); $p < 0.05$ (yellow); $p > 0.05$ (green)

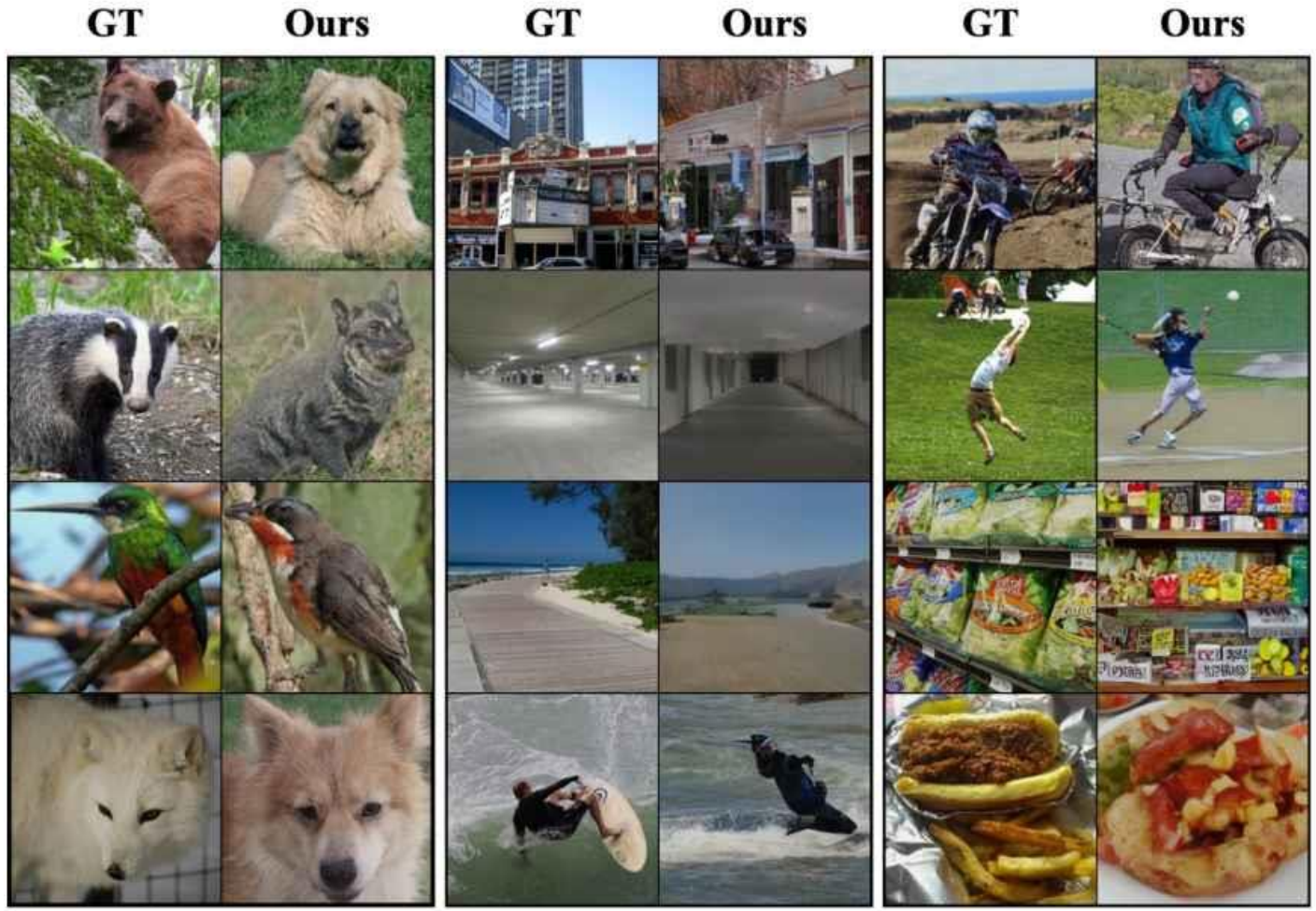


Figure 8. **Replication Dataset (BOLD5000).** It achieved similar quantitative results as the GOD dataset. 50-way top-1 identification accuracy: 34%; FID: 1.2 (Subject 1).

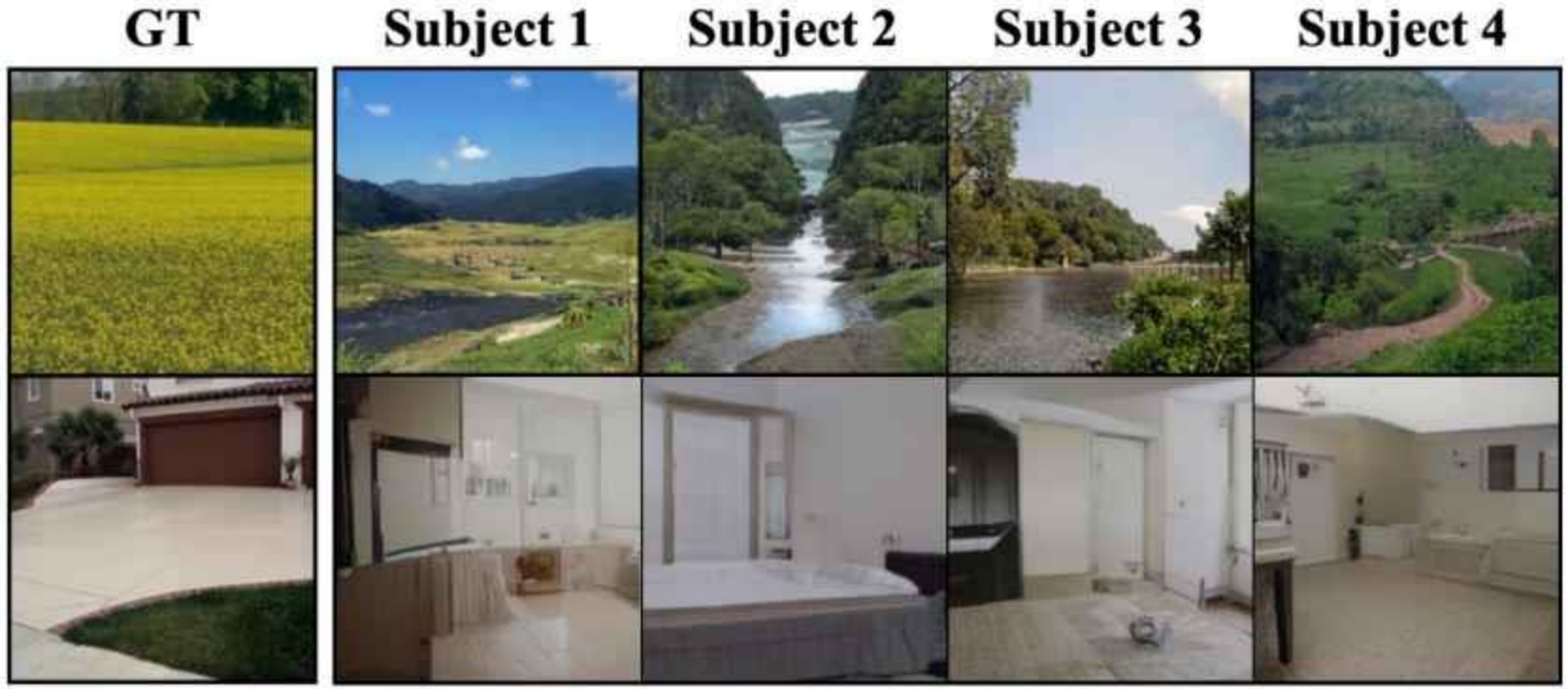


Figure 9. **Extra Features Decoded.** Imagery-related details can be decoded with our method. *e.g.* the river and blue sky were decoded with natural scenery stimulus (top row); similar interior decorating of indoor environments was decoded when a house was presented (bottom row).

BRAINLM: DATA

- **the UK Biobank (UKB)**

  task-based and resting-state functional MRI (fMRI) recordings plus medical records from over 40,000 subjects aged 40-69 years old. recordings were acquired on a Siemens 3T scanner at 0.735s temporal resolution

- **the Human Connectome Project (HCP)**

  1,002 high-quality fMRI recordings from healthy adults scanned at 0.72s resolution

- **Our model was trained on 80% of the UKB dataset (61,000 recordings) and evaluated on the held-out 20% and the full HCP dataset.**

- **Preprocessing**

  All recordings underwent standard preprocessing including motion correction, normalization, temporal filtering, and ICA denoising to prepare the data

  To extract parcel-wise time series, we parcellated the brain into 424 regions using the AAL-424 atlas [26]. This yielded 424-dimensional scan sequences sampled at 1 Hz.

Previous Method

Cui H et al. 2023. Brain Network Analysis with Graph Neural Network

the source code o BrainGB at https://github.com/HennyJie/BrainGB

Pre-training and Fine-tuning Transformers for fMRI Prediction Tasks

Code: https://github.com/GonyRosenman/TFF.

Community-Aware Transformer for Autism Prediction in fMRI Connectome

PTGB: Pre-Train Graph Neural Networks for Brain Network Analysis

The full implementation of this work is publicly available at https://github.com/Owen-Yang-18/BrainNN-PreTrain.

- **Three real-world brain network datasets**:

1) the Bipolar Disorder (BP) dataset,
2) the Human Immunodeciency Virus Infection (HIV) dataset,
3) the Parkinson's Progression Markers Initiative (PPMI) dataset

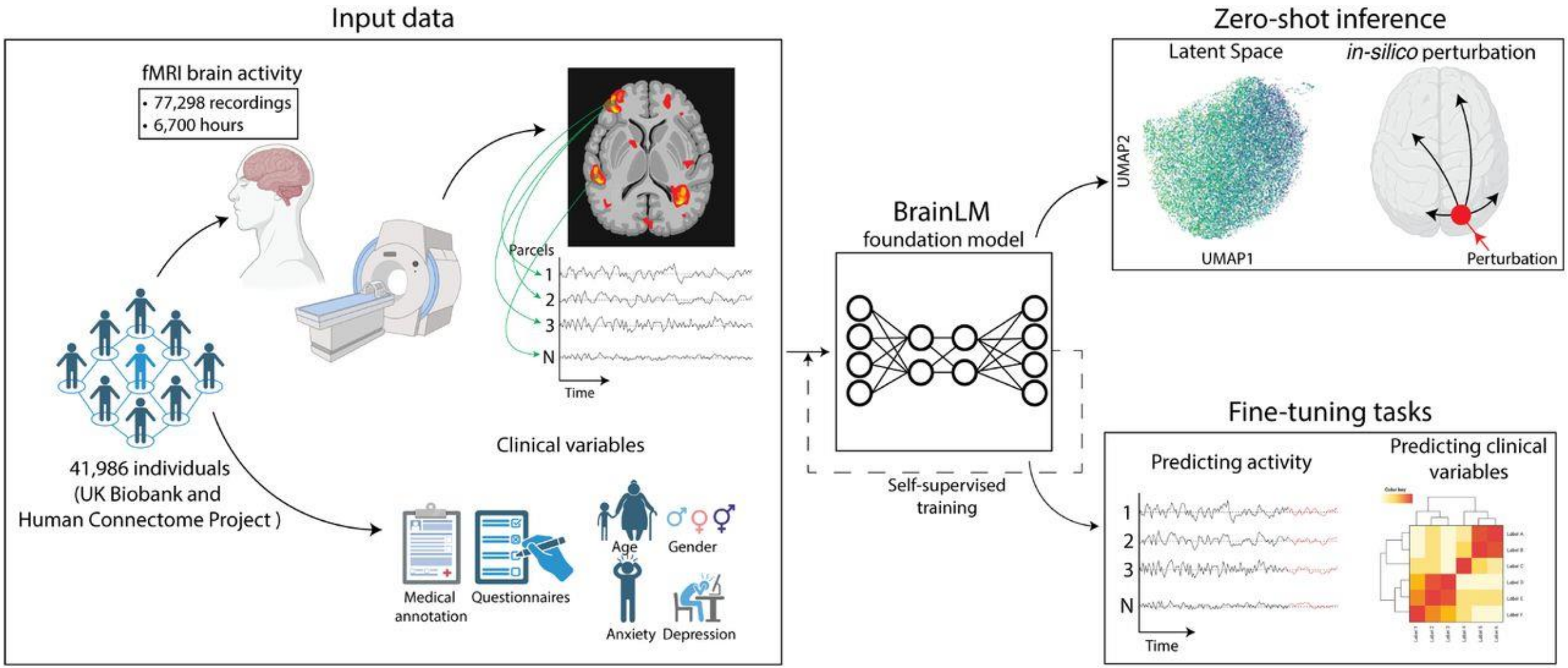
while the large-scale PPMI dataset is publicly available for authorized users

https://www.ppmi-info.org/

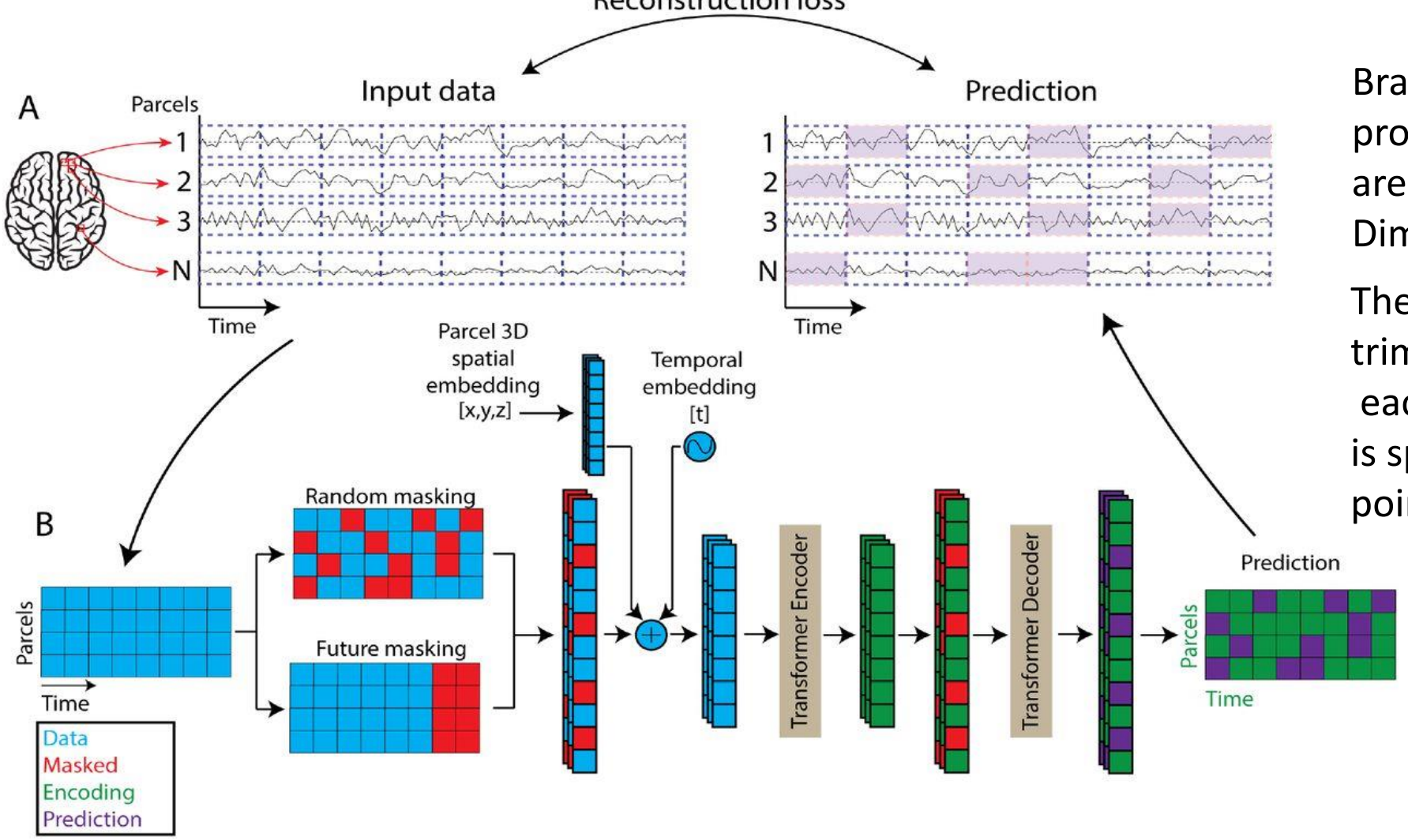
- **Data preprocessing pipelines provided by the opensource BrainGB platform**

https://braingb.us/

- **Our work is most closely related to recent efforts to apply masked autoencoders for unsupervised pretraining on fMRI data [13, 14] or other neural recording modalities, to learn substantially more powerful encodings of spatiotemporal fMRI patterns.**
- **BrainLM is the first model of its scale designed following the foundation model paradigm - pretrained on iverse unlabeled data and adaptable to various downstream applications via transfer learning.**
- **The code, model weights and training hyperparameters will be made publicly available**



Overview of the BrainLM framework. The model was pretrained on **6,700 hours of fMRI recordings** from 77,298 subjects via spatiotemporal masking and reconstruction. After pretraining, BrainLM supports diverse capabilities through fine-tuning and zero-shot inference. Fine-tuning tasks demonstrate prediction of future brain states and clinical variables from recordings. Zero-shot applications include **inferring functional brain networks** from attention weights and using a novel prompting technique **to simulate perturbation responses in silico**. This highlights BrainLM's versatility as a foundationmodel for fMRI analysis.



BrainLMarchitecture and training procedure. A) The fMRI recordings are compressed into 424 Dimensions (parcels)

The recordings are randomly trimmed to 200 time points. For each parcel, the temporal signal is split into patches of 20 time points each (blue dashed boxes).

The resulting 4240 patches are converted into tokens via a learnable linear projection.

Fromthe total number of tokens (blue), a subset is masked (red), either randomly or at future timepoints. We then add the learnable spatial and temporal embeddings to each token. These visible tokens (blue) are then processed by a series of Transformer blocks (Encoder). The input to the Decoder is the full set of tokens, consisting of encoded visible tokens (green) and masked tokens (red). The Decoder also consists of Transformer blocks and ultimately projects the tokens back to data space. Finally, we compute the reconstruction loss between the prediction (purple) and the original input data (blue).

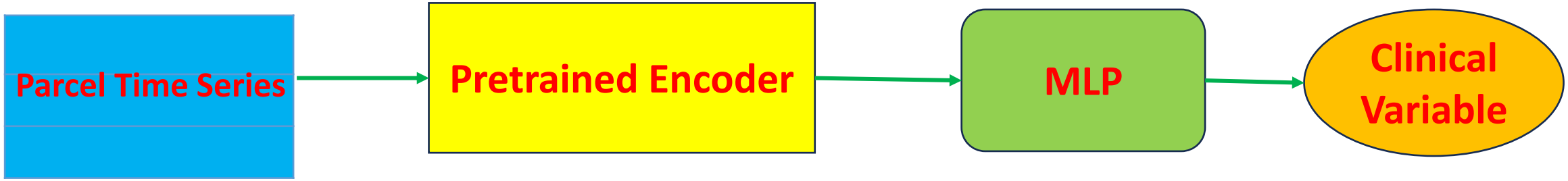
Training Procedure

- **For each fMRI recording, we sampled random200-timestep subsequences. The parcel time series were divided into segments of 20 timesteps, yielding 10 segments per subsequence. These were embedded into a 512-dimensional space and masked with a ratio of 20%, 50%, or 75%.**

  The unmasked segments were encoded via a Transformer encoder with 4 self-attention layers and 4 heads. This was decoded by a 2-layer Transformer to reconstruct all segments. We trained with batch size 512 and the Adam optimizer for 100 epochs, minimizing themean squared error between original and predicted embeddings

After pretraining on all sequences, the encoder can extract informative features capturing spatiotemporal brain activity patterns. We leverage the pre-trained encoder for downstream prediction and interpretation tasks.

Clinical Variable Prediction



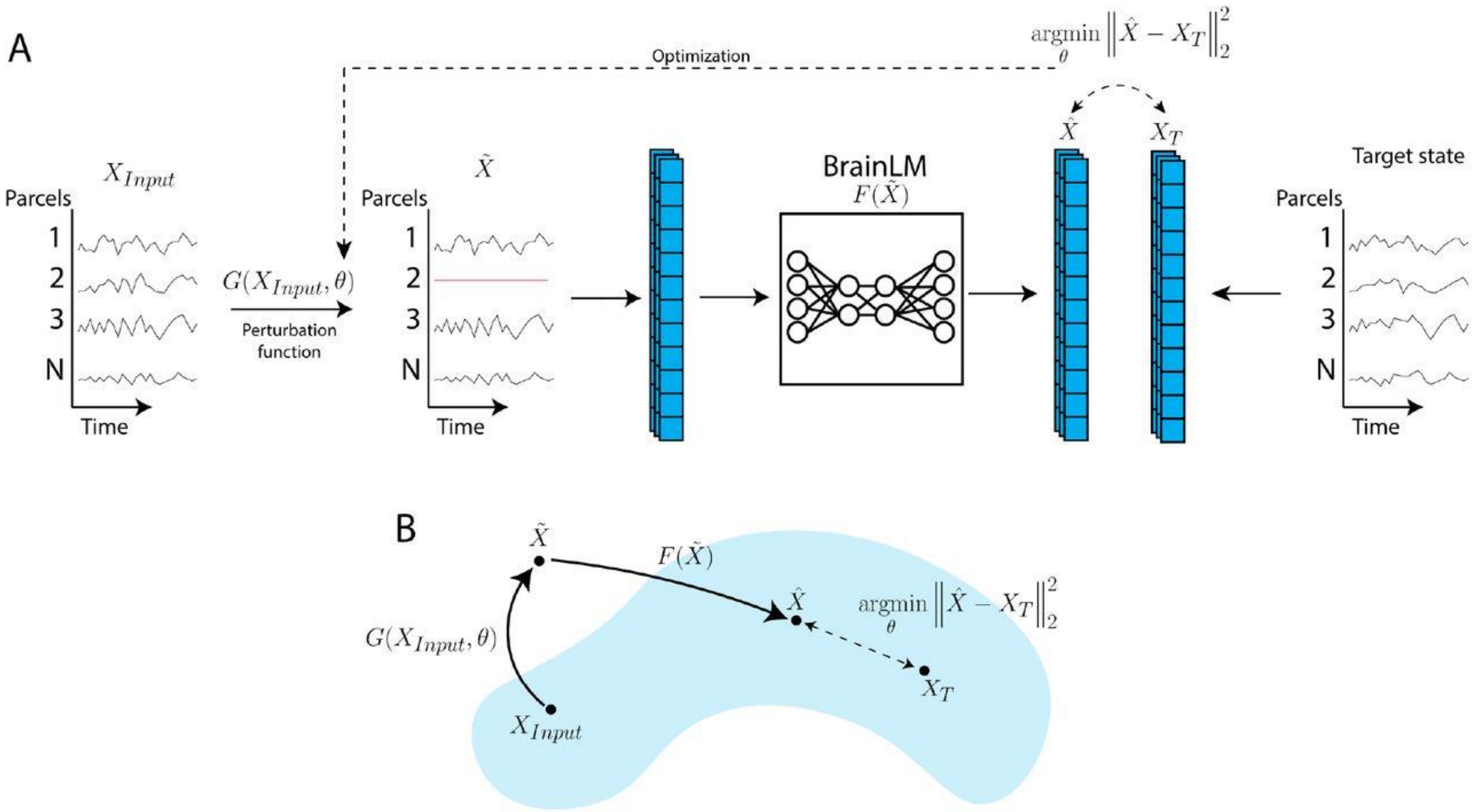
anxiety disorder scores
Age
Neuroticism score
disease

For age, we normalize by simply Z-scoring the Age values for all patients to a mean of 0 and unit variance.

For Neuroticism scores, we do min-max scaling to bring the distribution of scores into the range [0,1].

For Post Traumatic Stress Disorder (PCL) and General Anxiety Disorder (GAD7) scores, we first perform a log transformation tomake the values less exponentially distributed, and then performmix-max scaling to range.

In Silico Perturbation Simulation

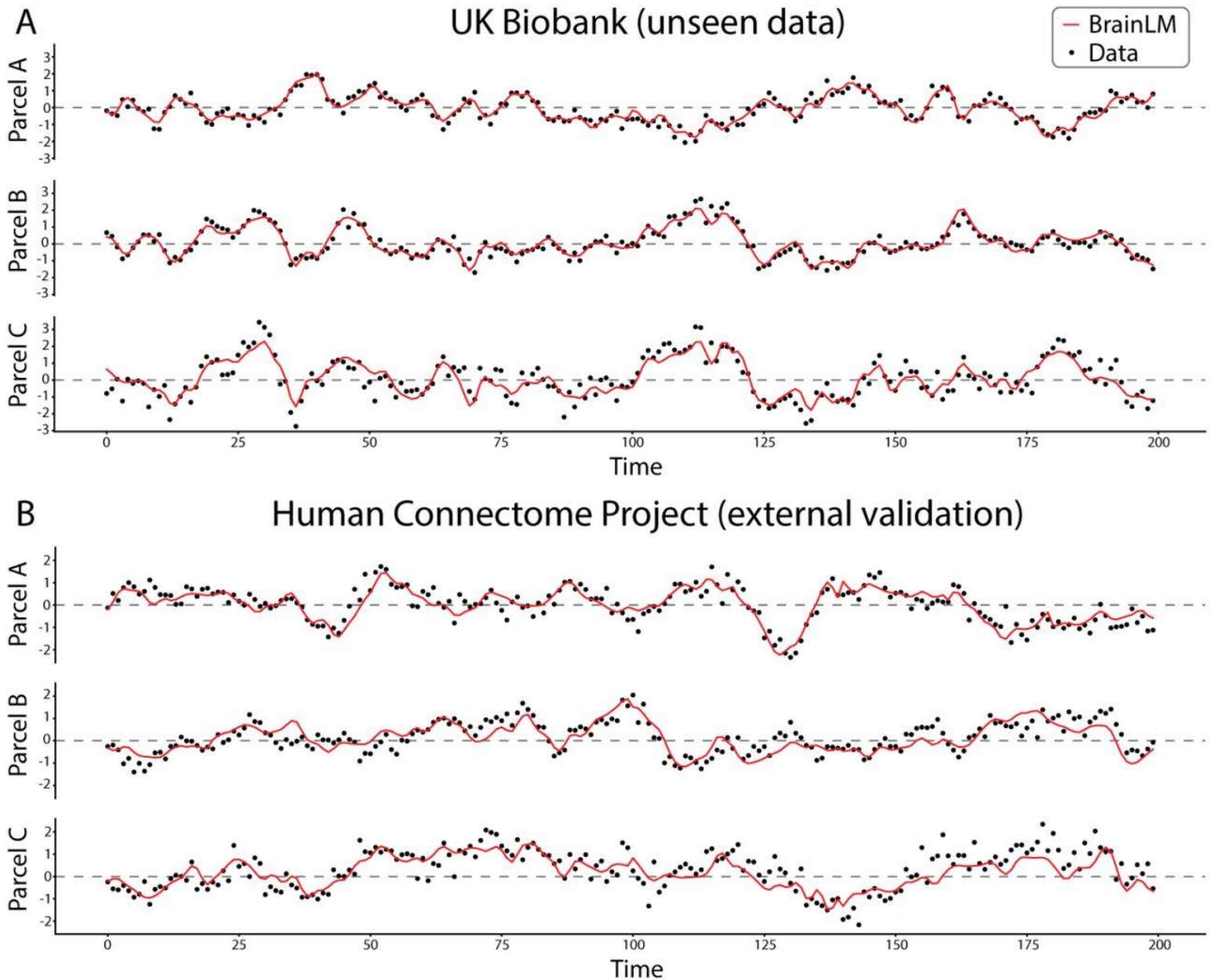


Results

Table 1: Performance comparison of latent space learned through self-supervised pretraining. Shown is the coefficient of determination (R^2) between predicted and ground truth data for masked patches across various configurations of masking ratio (MR) and training data size. Columns indicate models trained on 1% or 100% of the data and with 75% or 90% masking. Rows show different inference metrics, validated by masking 20%, 50%, or 75% on UK Biobank data (UKB) or Human Connectome Project (HCP) data. The top performing model was trained on 100% of the data with 75% making.

	MR=0.75		MR=0.90	
Data size	1%	100%	1%	100%
UKB (MR=0.2)	0.361	**0.402**	0.158	0.221
UKB (MR=0.5)	0.349	**0.389**	0.182	0.245
UKB (MR=0.75)	0.309	**0.343**	0.186	0.234
HCP (MR=0.2)	0.300	**0.316**	0.126	0.176

Model Generalization



Prediction of Clinical Variables

- A key advantage of foundation models is their ability to fine-tune on downstream tasks using the pretrained representations.
- The pretrained encoder was appended with an MLP head and fine-tuned to predict age, neuroticism, PTSD, and anxiety disorder scores. fine-tuning used a held-out subset of UKB subjects.
- 1) SVMs trained on raw fMRI data, and 2) SVMs trained on BrainLM's pretrained embeddings. Across all variables, (3) the fine-tuned BrainLM model

		Age	Post traumatic stress disorder (PCL)	General Anxiety disorder (GAD7)	Neuroticism
(1)	Raw data	2.0 ± 0.2219	0.034 ± 0.0027	0.172 ± 0.0066	0.160 ± 0.0137
(2)	BrainLM Pretrained	0.857 ± 0.1135	0.022 ± 0.0019	0.094 ± 0.0079	0.086 ± 0.0047
(3)	BrainLM fine-tuned	$\mathbf{0.485 \pm 0.0252}$	$\mathbf{0.018 \pm 0.0008}$	$\mathbf{0.074 \pm 0.0053}$	$\mathbf{0.072 \pm 0.0049}$

Prediction of Future Brain States

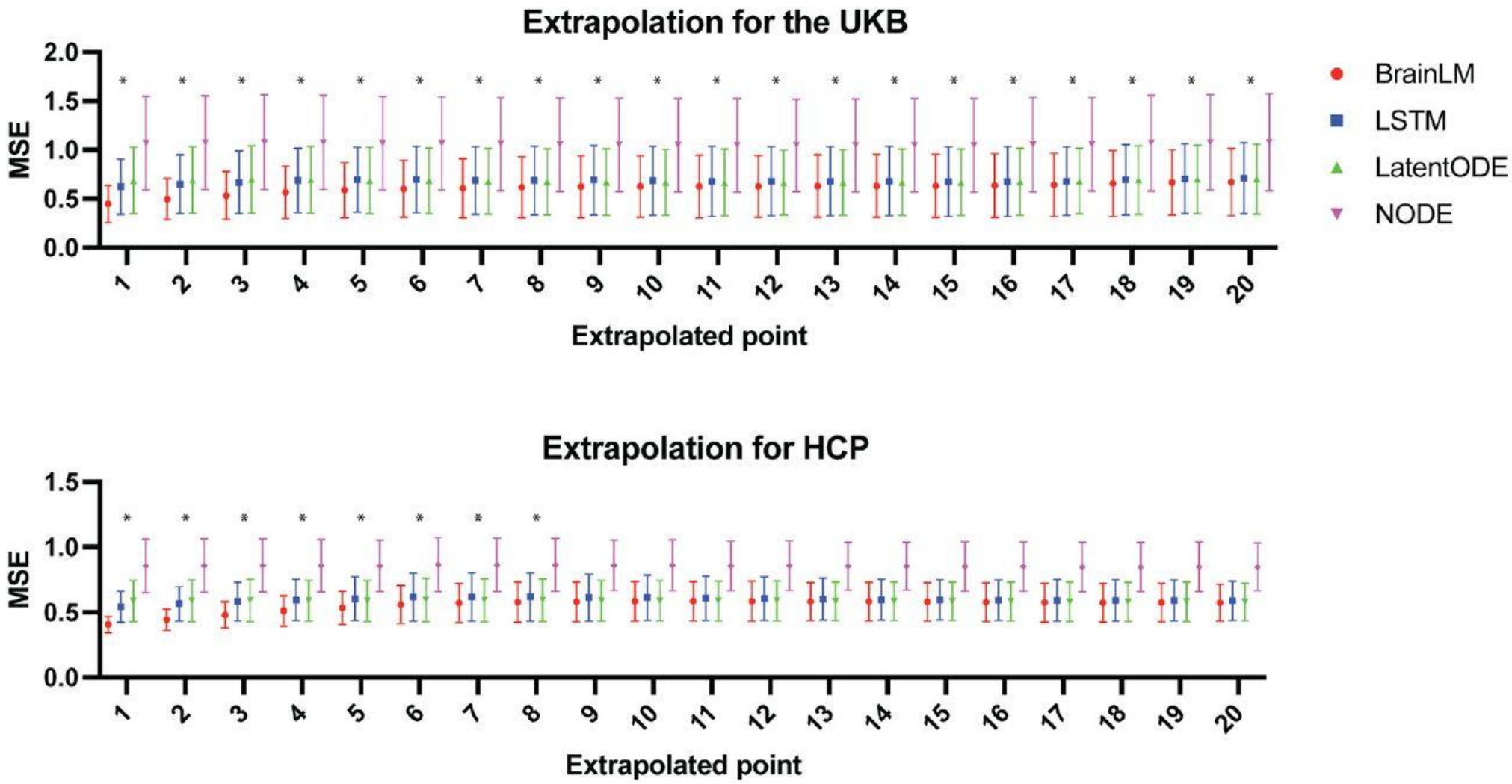
$$Y_{t+1}, \ldots, Y_{t+h} = f_\theta(Y_0, Y_1 \ldots, Y_t, X_0, X_1, \ldots, X_{t+h})$$

TimeGPT

Code: https://github.com/Spiderpig86/TimeGPT#-images

- **To evaluate whether BrainLM can capture spatiotemporal dynamics, we assessed its performance in extrapolating o future brain states**
- **During fine-tuning, BrainLM was given 180 timestep sequences and trained to forecast the subsequent 20 timesteps. We compared against baseline models including LSTMs, ODEnets, and a non-pretrained version of BrainLM**

Prediction of Future Brain States

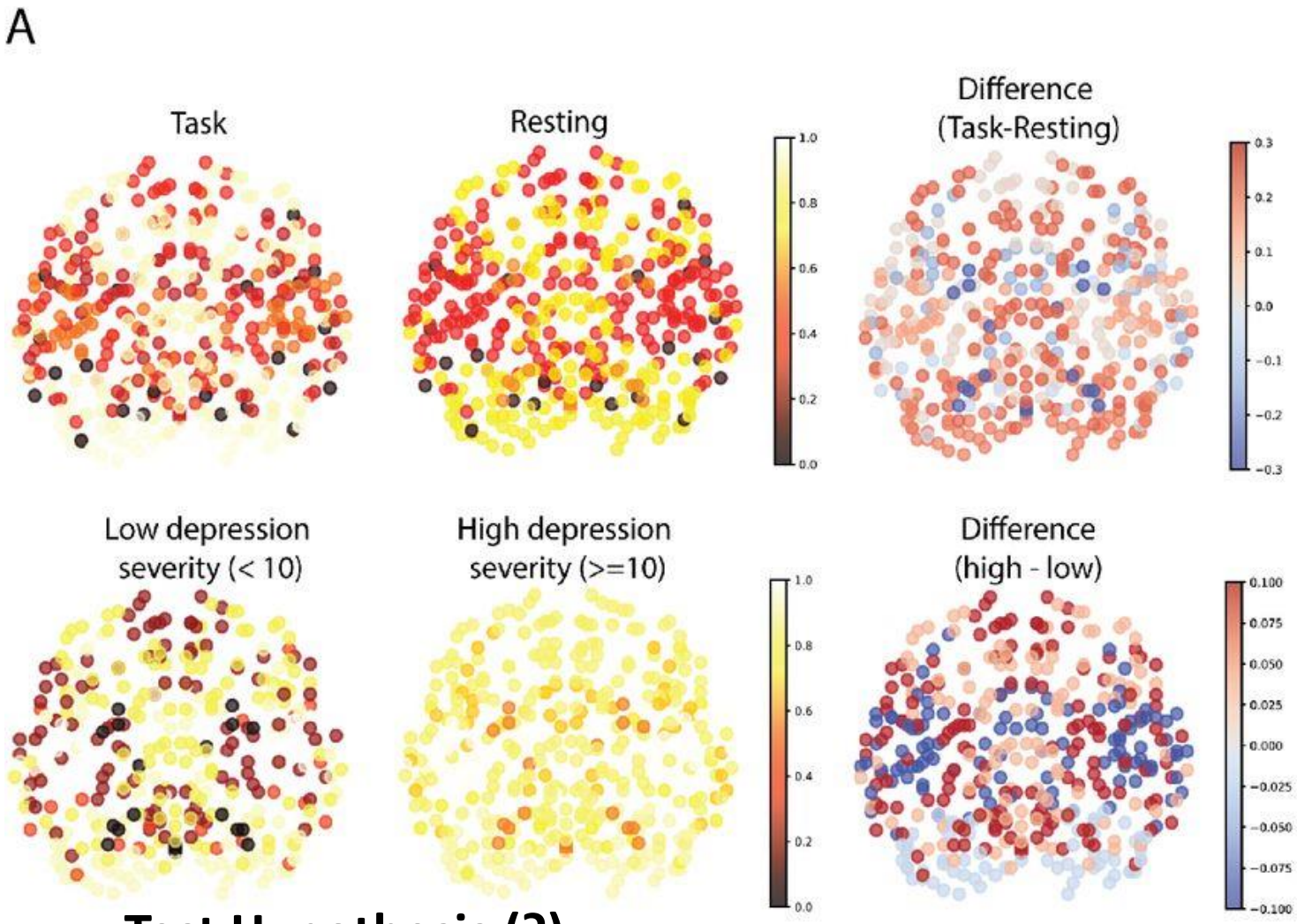


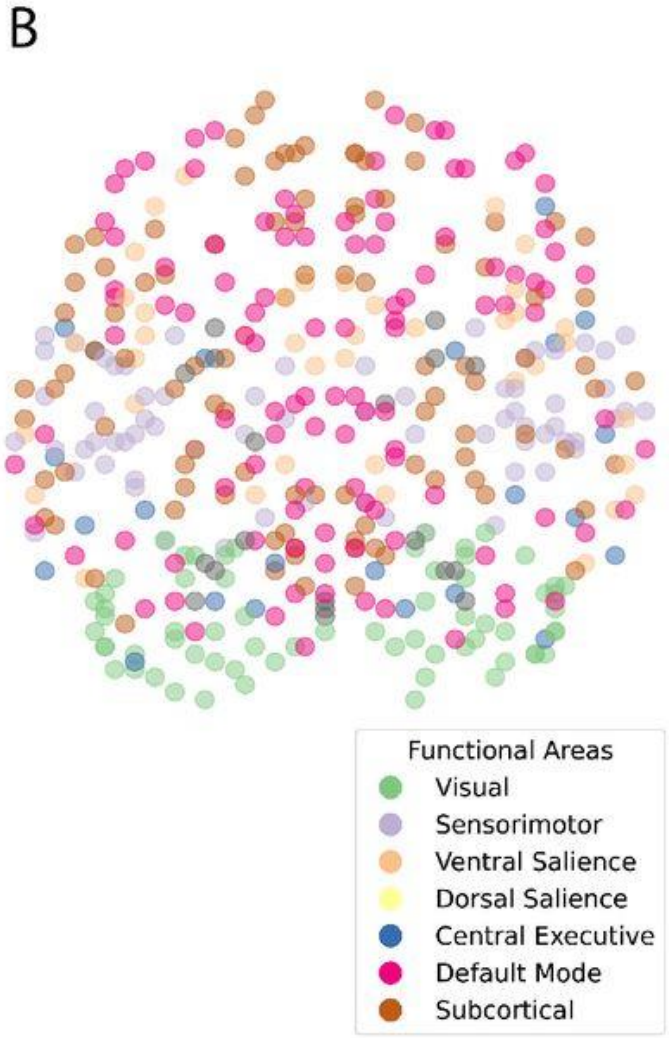
The time points for which BrainLMhas significantly ($p < 0.05$) lower error than the other methods are identified with "*".

Table 3: Quantitative evaluation of extrapolation performance. Models were tasked with forecasting parcel activity 40 timesteps beyond observed data from the UKB dataset. BrainLM shows the best performance across all metrics: higher (R2) and Pearson correlation coefficients (R), and lower mean squared error (MSE) between predicted and true future states.

	UKB			HCP		
	R^2	R	MSE	R^2	R	MSE
BrainLM (fine-tuned)	**0.086**	**0.280**	**0.648**	**0.028**	**0.185**	**0.568**
BrainLM (w/o pre-training)	0.012	0.112	0.695	0.007	0.090	0.583
LSTM	-0.001	0.151	0.704	-0.020	0.049	0.598
Neural ODE	-0.577	0.001	1.083	-0.469	2.010e-4	0.857
Latent ODE	0.001	0.023	0.703	-0.003	-2.026e-4	0.588

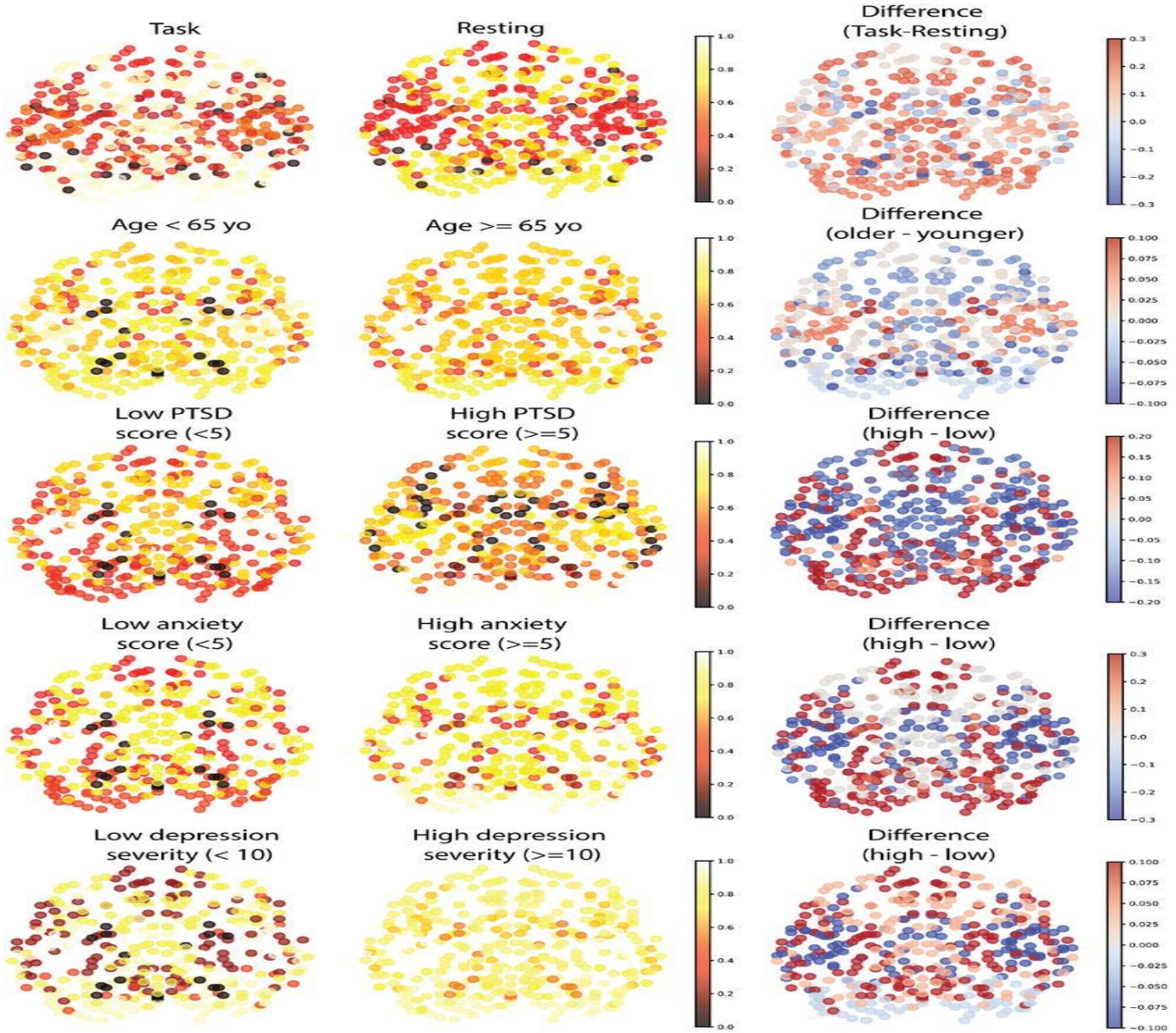
Interpretability via Attention Analysis



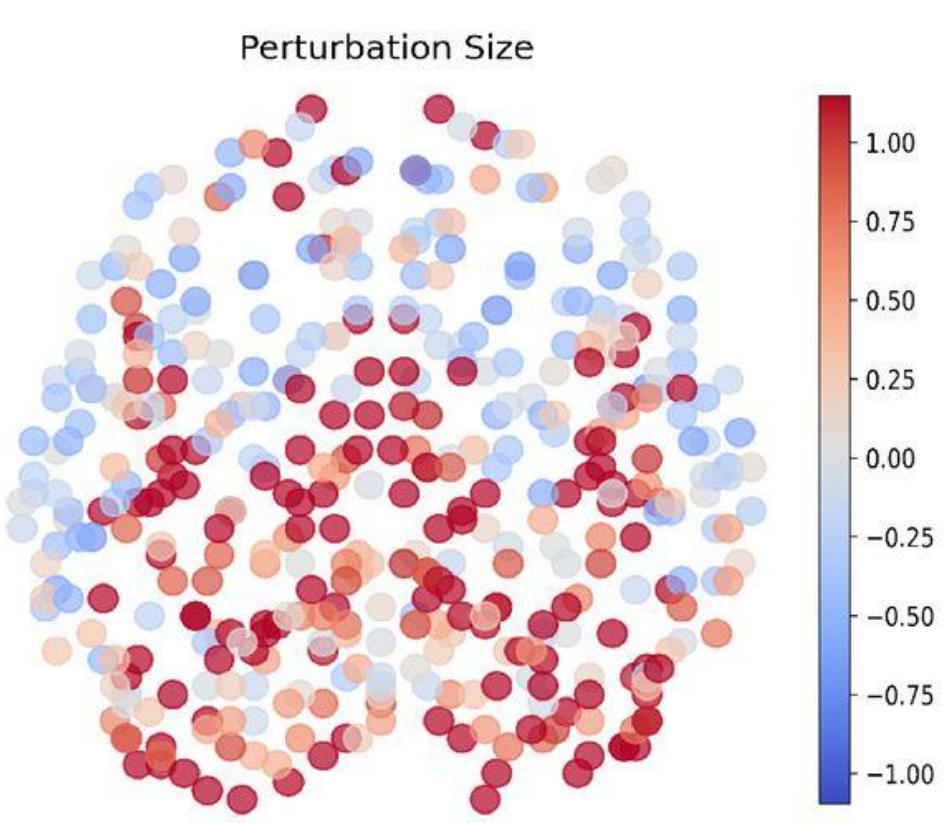


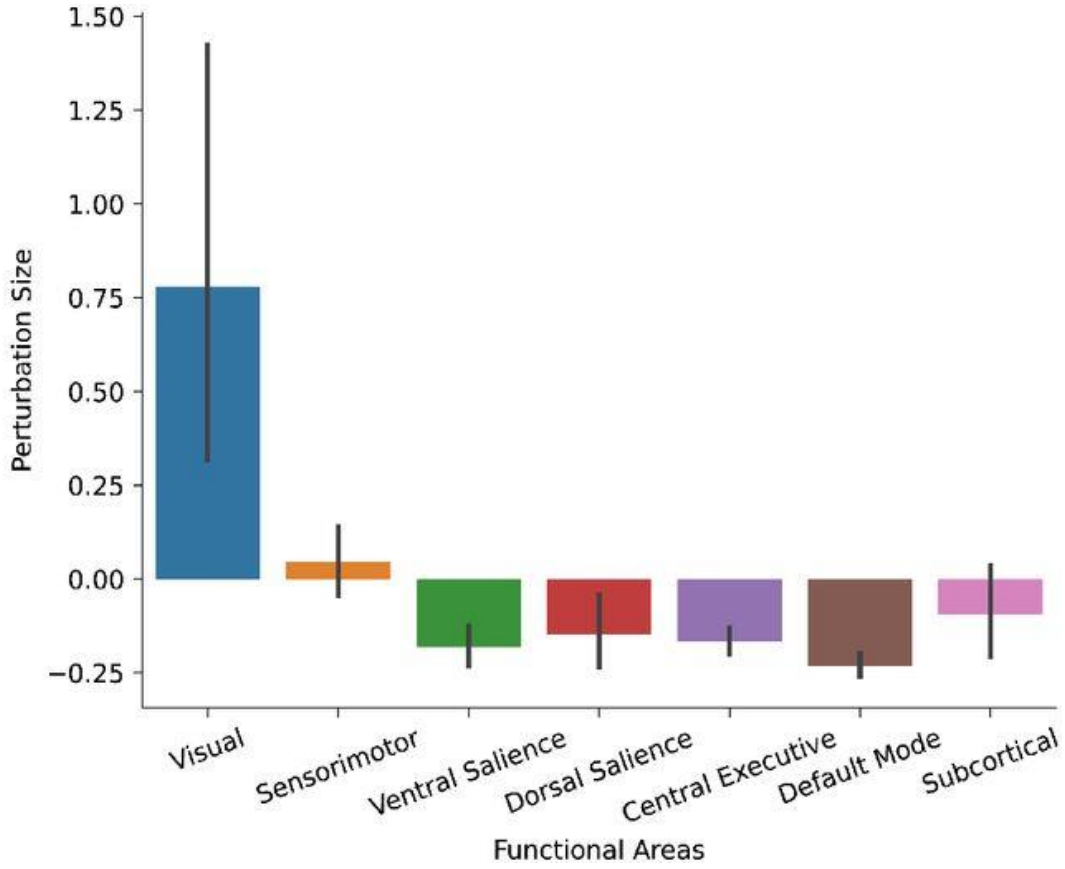
Test Hypothesis (?)

Interpretability via Attention Analysis



In silico Perturbation Analysis Reveals Functional Connectivity





In silico perturbation of resting state to match task-based recordings reveals functional changes. The average magnitude of optimized perturbations to make resting state CLS tokens match target task CLS tokens. We find that the region with the largest predicted perturbation is the visual cortex, in line with expected functional changes between resting state and task-based recordings.

Functional Network Prediction

We evaluated BrainLM's ability to segment parcels into intrinsic functional brain networks directly from fMRI activity patterns, without any network-based supervision. Parcels were categorized into 7 functional groups as defined in prior cortical parcellation work [cite]. The groups corresponded to visual, somatomotor, dorsal attention, ventral attention, limbic, frontoparietal, and default mode networks. On a held-out set of 1,000 UKB recordings, we compared different methods for classifying parcels into these 7 networks:

The classifiers were trained on 80% of the parcels from each recording and evaluated on the remaining 20%.

	Accuracy (%)
BrainLM (attention weights)	**58.8**
Raw Data	39.2
Variational Autoencoder	49.4
Graph Convolutional Network	25.9

Projects

- **Data Sets:**

MinD-Vis: Sparse Masked Brain Modeling with Double-Conditioned Latent Diffusion Model for Human Vision Decoding .

Human Connectome Project (HCP) 1200 Subject Release
[55];
Generic Object Decoding Dataset (GOD) [21]; and Brain,
Object, Landscape Dataset (BOLD5000)

the UK Biobank (UKB)

Above Pretraining

PTGB: Pre-Train Graph Neural Networks for Brain Network Analysis

Three real-world brain network datasets:
the Bipolar Disorder (BP) dataset,
2) the Human Immunodeciency Virus Infection (HIV) dataset,
3) the Parkinson's Progression Markers Initiative (PPMI) dataset

while the large-scale PPMI dataset is publicly available for authorized users

https://www.ppmi-info.org/

Repeat Analysis in Brain LM