CBrain: Cross-Modal Learning for Brain Vigilance Detection in Resting-State fMRI

Chang Li[1], Yamin Li[1], Haatef Pourmotabbed[1], Shengchao Zhang[2], Jorge A. Salas[1], Sarah E. Goodale[1,3], Roza G. Bayrak[1], and Catie Chang[1]

[1] Vanderbilt University, Nashville, TN, USA
chang.li@vanderbilt.edu
[2] Rhode Island Hospital (Brown University Health), RI, USA
[3] Vanderbilt University Medical Center, TN, USA

Abstract. Detecting human vigilance states (e.g., natural shifts between alertness and drowsiness) from functional magnetic resonance imaging (fMRI) data can provide novel insight into the whole-brain patterns underlying these critical states. Moreover, as a person's vigilance levels are closely tied to their behavior and brain activity, vigilance state can strongly influence the results of fMRI studies. Therefore, the ability to annotate fMRI scans with vigilance information can also enable clearer and more robust results in fMRI research. However, well-established vigilance indicators are derived from other modalities such as behavioral responses, electroencephalography (EEG), and pupillometry, which are not typically available in fMRI data collection. While previous works indicate the promise of distinguishing vigilance states from fMRI alone, EEG data can provide reliable vigilance indicators that complement and augment fMRI domain information. Here, we propose **CBrain**: **C**ross-modal learning for **Brain** vigilance detection in resting-state fMRI. Our model transfers EEG vigilance information into an fMRI latent space in training, and predicts human vigilance states using only fMRI data in testing, addressing the need for external vigilance indicators. Experimental results demonstrate CBrain's ability to predict vigilance states across different individuals at a granularity of 10-fMRI-frames with an **81.07%** $mF1$ score on a test set of unseen subjects. Additionally, our generalization experiments highlight the model's potential to estimate vigilance in an unseen task and in resting-state fMRI scans collected with a different scanner at a different site. Source code: https://github.com/neurdylab/CBrain.

Keywords: fMRI · EEG · brain patterns · cross-modal learning

1 Introduction

Functional magnetic resonance imaging (fMRI) data provides whole-brain blood-oxygen-level-dependent (BOLD) signals at millimeter-scale resolution. During fMRI scans, subjects often experience natural shifts between alertness and drowsiness, also known as fluctuations in *vigilance*. Deriving brain-wide patterns associated with vigilance provides a window into brain activity associated with this

critical brain state[3, 13]. Further, vigilance state fluctuations strongly influence subjects' behavior and brain activity[13]. As a result, modeling the effects of vigilance on fMRI data, and providing accurate vigilance state annotations, can facilitate fMRI research and enable a more robust interpretation results across a broad range of studies.

Well-established vigilance indicators can be derived from measures such as EEG and pupillometry. However, collecting these additional data requires specialized hardware, additional setup time, and expertise in acquiring and denoising these signals. Resting-state fMRI scans also lack behavioral measures, which can also convey vigilance information. The lack of ground-truth vigilance indicators in most fMRI datasets motivates the development of methods to infer vigilance fluctuations directly from fMRI data. Previous work has demonstrated the feasibility of detecting vigilance from fMRI [8, 31, 23]. However, these methods are either limited in temporal resolution, predicting vigilance states from functional connectivity windows of $\geq$ 40s[23, 26] or they have high temporal resolution but are limited in their ability to quantify corresponding states and state transitions across scans (as opposed to relative variations within scans)[8], or are unsupervised methods that need additional information to label the discovered states[31]. fMRI foundation models[1, 2, 6, 29] are demonstrating promise in applications such as fMRI signal reconstruction and disease modeling. However, to our knowledge, such models have not yet been leveraged for predicting vigilance states from fMRI.

EEG is highly complementary to fMRI: fMRI has high spatial resolution but lower temporal resolution, and measures a blood-oxygen response; while EEG has lower spatial resolution but high temporal resolution, and measures neuroelectrical activity. Moreover, EEG data is rich with vigilance-related information [19]. Combining fMRI and EEG may thus provide improved spatiotemporal features, and allows for incorporating EEG vigilance information into latent representations of fMRI data. We hypothesize that leveraging EEG to train an fMRI vigilance-detection model may increase a model's ability to discern vigilance states. Currently, fMRI vigilance labeling strategies are not yet established as ground truth. By training fMRI models to predict EEG-derived labels, we address an open challenge: enabling vigilance detection from fMRI alone.

We propose **CBrain**: **C**ross-modal learning for **Brain** vigilance detection, a cross-modal architecture that leverages EEG to enhance fMRI-based vigilance detection. We train our model on simultaneous EEG-fMRI data, transferring vigilance-related cross-modal EEG knowledge into the fMRI latent space. During testing, CBrain uses only 10-frame fMRI data patches as input and achieves an **81.07%** $mF1$ score in predicting vigilance states for unseen test-set subjects, addressing the challenge of distinguishing brain states in fMRI without relying on other data modalities. Our experiments on another EEG-fMRI dataset, acquired on a different scanner, demonstrate our model's ability to generalize to both resting-state and task scans. These results show that CBrain has the potential to enrich existing fMRI scans, including those in large public datasets that lack vigilance measures[24, 17], with new information about vigilance state.

2 Method

Our model is trained on 10-frame fMRI data patches and corresponding EEG data patches. We perform fMRI intra-modal learning, fMRI-EEG cross-modal learning, and prediction head training in one stage. In testing, our model predicts vigilance states based only on the fMRI data input, as shown in Fig.1.

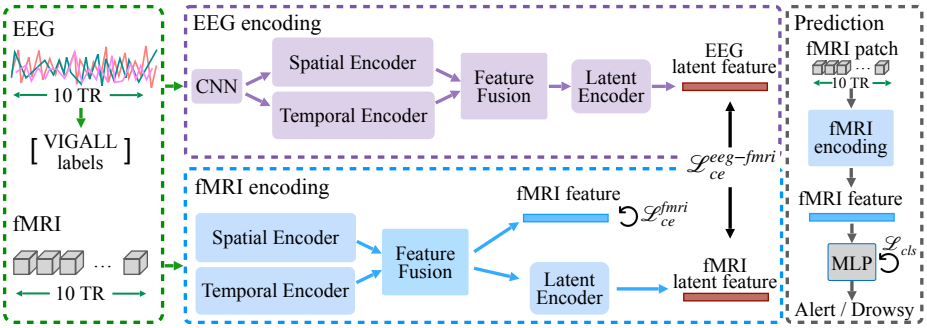

Fig. 1. CBrain's pipeline. We train our model on paired 10-frame fMRI data and EEG data segments. In testing, our model takes only fMRI data as input.

Obtaining Patch-wise Vigilance Ground-truth: In a simultaneous EEG-fMRI dataset, for each fMRI scan X with T frames, we calculate frame-wise vigilance score $V_{1..T}$ by applying Vigilance Algorithm Leipzig[19] on its paired EEG data Y, and convert the integer values to range (-1, 0, 1) following [21]. We sum frame-wise labels for every 10-frame fMRI patch and assign the ground truth as alert if the sum exceeds -5, and drowsy otherwise.

Cross-modal Contrastive Learning: For each 10-frame fMRI data segment x, consisting of data from 64 regional and 2 global fMRI time courses (described in Section 3.1), we first extract its spatial features f_x^s and temporal features f_x^t by performing attention on the spatial and temporal axes using spatial and temporal transformer-based[25] encoders respectively. Then we fuse f_x^s and f_x^t with a given feature fusion ratio to obtain fMRI-domain spatial-temporal features f_x by:

$$f_x = ratio * f_x^s + (1 - ratio) * f_x^t \tag{1}$$

For the fMRI segment's corresponding 26-channel EEG data patch y, we apply a 1D CNN downsampling layer, then harvest EEG spatial features f_y^s and temporal features f_y^t from EEG spatial and temporal transformer encoders in the same manner as in fMRI domain. We fuse these EEG spatial and temporal features using the same feature ratio as fMRI features. We map fMRI features f_x and EEG features f_y to a common latent space using our latent encoding modules, each consisting of two blocks of a linear layer followed by one ReLU activation, obtaining f_x^{latent} and f_y^{latent}. Then, we apply contrastive learning on the fMRI features and cross-modal latent space, allowing the model to capture intrinsic

vigilance-state differences and to bridge EEG brain-state-related features into the fMRI domain. In the fMRI domain, we perform contrastive learning on fMRI fused features f_x to increase the fMRI features' discriminability. For cross-modal learning, we apply contrastive loss on the fMRI latent features f_x^{latent} and corresponding EEG latent features f_y^{latent}, bringing EEG brain-state information into a shared latent space with fMRI features. We use contrastive learning loss[4, 15], defined for a batch of features f with m samples with temperature τ as:

$$\mathcal{L}_{ce}(f_{1..m}) = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{\sum_{k=1}^{m_i^{pos}} e^{f_i^\top f_k/\tau}}{\sum_{j=1}^{m} e^{f_i^\top f_j/\tau}}, \tag{2}$$

where m_i^{pos} denotes the positive samples for feature f_i. For feature f_i, positive samples are features with the same ground truth label as f_i; negative samples are features with different labels. In practice, we only integrate the contrastive learning loss for the first sample in the batch. The fMRI-modal contrastive loss $\mathcal{L}_{ce}^{fmri}$ and cross-modal contrastive loss $\mathcal{L}_{ce}^{eeg-fmri}$ are as follows:

$$\mathcal{L}_{ce}^{fmri} = \mathcal{L}_{ce}(f_{x_{1..m}}) \tag{3}$$

$$\mathcal{L}_{ce}^{eeg-fmri} = \mathcal{L}_{ce}(f_{x_{1..m}}^{latent}) + \mathcal{L}_{ce}(f_{y_{1..m}}^{latent}) + \mathcal{L}_{ce}(f_{x_{1..m}}^{latent}, f_{y_{1..m}}^{latent}) \tag{4}$$

Vigilance State Prediction: Given a 10-frame segment of fMRI data from an unseen subject, we first extract its spatial-temporal features without mapping it to the latent space, and feed it to a classification MLP head with three hidden layers, each comprising a linear layer, batch normalization, and LeakyReLU activation. The output layer is a linear layer without normalization. The MLP head is trained using cross-entropy loss $\mathcal{L}_{cls}$. The overall training loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{ce}^{fmri} + \mathcal{L}_{ce}^{eeg-fmri} \tag{5}$$

Model Training: We train our model in an end-to-end manner, with the feature fusion ratio as 0.5, contrastive learning temperature τ as 0.1, the weights of $\mathcal{L}_{cls}$, $\mathcal{L}_{ce}^{fmri}$, $\mathcal{L}_{ce}^{eeg-fmri}$ as 0.5, 0.1, and 0.1. We train our model for 50 epochs using AdamW optimizer, a learning rate of 7e-4, a weight decay of 0.1, and a batch size of 32 on a single NVIDIA RTX 6000 GPU. Training on the full dataset takes less than one hour. We adopt a linear warm up from 1e-6 for 20 epochs. We develop our code based on the frameworks in[18, 15] to perform cross-modal contrastive learning for vigilance detection. CBrain's theoretical complexity is as follows: encoder[25]: $O(n^2 * d)$, MLP: $O(m * d_{1..n})$, contrastive learning[4]: $O(m^2)$ (n: sequence length, d: feature dim, m: batch size).

3 Experiments

3.1 Training and External Validation Datasets

For model training, we used a simultaneous EEG-fMRI dataset comprising 29 resting-state scans from 22 healthy subjects. Subjects provided written informed

consent, and protocols were approved by the Institutional Review Board. Functional MRI data was acquired using a multi-echo EPI sequence (3T scanner, TR=2100ms). 32-channel EEG data was collected at 5KHz, synchronized to the scanner's 10MHz clock. Briefly, for fMRI data preprocessing, slice-timing correction, motion coregistration, multi-echo ICA denoising, alignment to MNI152 space, and 3mm spatial smoothing were performed. Using the Dictionary of Functional Modes atlas[5], 64 regions of interest (ROIs) were extracted. Additionally, 1st through 4th-order polynomial trends and 6 motion parameters were regressed out of the data. The global signal with and without motion parameters were calculated and added to the input as two additional channels. We lagged the fMRI data by 2TRs (4.2 sec) with respect to the EEG data to account for hemodynamic response delay. EEG data was corrected for MRI and ballistocardiogram artifacts and downsampled to 250 Hz. We use 26 channels in training(EMG/ECG channels excluded). We obtain fMRI (and paired EEG) data segments with a sliding time window of 10 fMRI frames and a step size of 5 fMRI frames. Please refer to [12] for the dataset and preprocessing details. The training set includes 80% of the subjects (17 subjects, 23 scans), and the test set comprises 20% (5 subjects, 6 scans), with no subject overlap between training and testing splits. Noisy EEG segments were identified according to [7] and were excluded from training and testing. Another EEG-fMRI dataset, acquired at a different site and with different fMRI paradigms, was used for external validation (with no data used in training)[8]. This dataset comprises scans from 14 healthy subjects (3T, TR=2100ms), who collectively underwent 16 eye-closed-rest scans and 12 scans collected during the delivery of intermittent auditory tones (also with eyes closed). We preprocessed this dataset the same way as the dataset described above.

3.2 Main Results

The main results, and comparisons with baselines trained with fMRI data only, are shown in Table 1. Our baselines include: fMRI classification models with strong performance in [20], timeseries models tailored for brain imaging tasks in [28], an fMRI foundation model[2] and the hierarchical SVM with high accuracy in a sleep staging task[23]. **CBrain** achieves a macro mean-F1 ($mF1$) score of 81.07% in the testing set, surpassing all baselines. The comparison between full-CBrain with fMRI-only CBrain and fMRI-only baselines shows that incorporating EEG data boosts CBrain's performance (3.22%), supporting our key contribution that EEG knowledge enhances fMRI-based brain state detection. This demonstrates the power of integrating complementary modalities and the potential of cross-modal supervision as an emerging paradigm in this field.

CBrain also yields consistently high classification $mF1$ in the eye-closed-rest (ecr) external validation data. CBrain trained with fMRI data alone attains a competitive performance of 77.85% $mF1$ in the testing set and 78.44% $mF1$ in the ecr external validation set, only inferior to the full CBrain model and the attention MLP[20], demonstrating the effectiveness of our intra-model contrastive learning in differentiating vigilance states. We acknowledge that meanMLP[20],

Table 1. CBrain's performance on testing set and eye-closed-rest external validation scans, compared to baselines, with (subject-wise mean ± std) for best performing models. $F1_d$: $F1_{drowsy}$, $F1_a$: $F1_{alert}$. **Bold with underline:** best performance. Underline: second-best performance. Random guess: output random predictions on brain states.

Methods	Testing Set			External Validation (ecr)		
	$F1_d$	$F1_a$	$mF1$	$F1_d$	$F1_a$	$mF1$
random guess	55.71	42.20	48.95(52.2 ± 9.7)	50.32	46.31	48.31(57.0 ± 9.6)
BrainLM[2]	64.82	37.25	51.04	59.72	37.13	48.43
BolT[1]	73.63	51.52	62.58	65.10	61.39	63.25
Medformer[28]	78.40	50.88	64.64	72.95	56.41	64.68
Nonformer[14]	76.86	55.40	66.13	76.40	69.94	73.17
BrainNetCNN[10]	78.98	58.77	68.87	72.69	66.70	69.69
SVM[23]	82.40	58.18	70.29	79.61	70.18	74.89
meanLSTM[20]	81.05	65.51	73.28(68.5 ± 16.1)	78.52	77.13	77.82(77.2 ± 15.0)
attnMLP[20]	85.39	70.53	<u>77.96</u>(70.4 ± 17.2)	79.10	75.85	77.48(78.2 ± 14.3)
meanMLP[20]	84.59	69.57	77.08(70.2 ± 17.0)	83.41	81.20	**<u>82.31</u>**(81.0 ± 13.3)
CBrain (fMRI-only)	84.50	71.21	77.85(71.2 ± 15.0)	79.54	77.35	78.44(77.3 ± 17.2)
CBrain	86.20	75.95	**<u>81.07</u>**(75.5 ± 12.2)	79.63	78.01	<u>78.82</u>(78.6 ± 15.7)

a very recent model, outperforms CBrain in the external validation set; the reason for this difference merits further investigation. We speculate that the slight drop in CBrain's performance on the external validation data might be caused by differences in dataset statistics, and future work might examine potential improvements through 1) data normalization strategies, and 2) encoding demographic and hardware information to handle inter-subject and inter-site variability.

3.3 Ablation Studies and Qualitative Analysis

Comparison with EEG Foundation Models: In Table 2, we compare our EEG encoder with state-of-the-art EEG foundation models that benefit from pre-training techniques to extract high-quality EEG features[27, 30, 11, 9] compared in [27]. We finetune the released checkpoints on our training set to extract EEG features and use 1D CNN for downsampling before applying our latent encoder. Our original model has the best $mF1$ score, demonstrating that by fusing spatial-temporal features, we can extract high-quality EEG features that better assist fMRI models in improving the fMRI latent space's discriminability. Leveraging pre-trained EEG features leads to high performance, as utilizing BIOT[30], BENDR[11], and LaBraM[9] reached an $mF1$ that exceeded all fMRI-only models in the testing set. This reveals that EEG knowledge can enhance the fMRI model's understanding of brain states.

Training Loss: CBrain performs best with all loss components as shown in Table 3. These results further demonstrate that complementary EEG knowledge can improve the fMRI model's ability to differentiate brain patterns.

Generalization: CBrain has a strong performance on the unseen external validation dataset collected at a different site and scanner, suggesting its gener-

alizability to unseen data (Table 4). A comparison between CBrain trained on fMRI data alone versus EEG-fMRI data shows that integrating EEG improves its robustness on external datasets beyond resting-state scans.

Qualitative Analysis: Fig.2 shows the predictions of the full CBrain model, together with ground-truth vigilance states and the corresponding EEG spectrograms. In Fig.3, we visualize the fMRI feature space for the testing set and the external validation eye-closed-rest scans. UMAP[16] scatter plots (first column) show that our fMRI encoders can successfully extract fMRI features with strong brain-state discriminability. We also visualize the average alert and drowsy fMRI features in the test and external validation (ecr) sets and project them onto the brain surface map. In line with prior work[31], the drowsy patterns show greater sensorimotor and visual activity. These results demonstrate that CBrain can generate consistent fMRI embeddings for different brain states.

Table 2. Comparison of EEG encoders on the testing set (with subject-wise mean ± std). **Bold:** best performance.

Method	$F1_{drowsy}$	$F1_{alert}$	$mF1$	mAC
CBrain w/ EEGPT[27]	83.89	70.18	77.03(72.6 ± 13.6)	79.08
CBrain w/ BIOT[30]	86.08	72.60	79.34(71.9 ± 15.9)	81.54
CBrain w/ BENDR[11]	85.99	75.42	80.71(75.6 ± 14.4)	82.15
CBrain w/ LaBraM[9]	85.61	75.98	80.79(75.3 ± 14.0)	82.00
CBrain	86.20	75.95	**81.07**(75.5 ± 12.2)	**82.46**

Table 3. Ablations on training loss with training weights. **Bold:** best performance.

Method	$\mathcal{L}_{cls}$	$\mathcal{L}_{ce}^{fmri}$	$\mathcal{L}_{ce}^{eeg-fmri}$	$F1_{drowsy}$	$F1_{alert}$	$mF1$	mAC
CBrain	0.5	-	-	81.86	67.10	74.48	76.62
	0.5	0.1	-	83.21	69.96	76.59	78.46
	0.5	0.1	0.1	86.20	75.95	**81.07**	**82.46**

Table 4. Generalization experiments on external validation dataset in eye-closed-rest and eye-closed-task scans with (subject-wise mean ± std). $F1_d$: $F1_{drowsy}$, $F1_a$: $F1_{alert}$. **Bold:** best performance.

Methods	**Eye-closed-rest**			**Eye-closed-task**		
	$F1_d$	$F1_a$	$mF1$	$F1_d$	$F1_a$	$mF1$
CBrain (fMRI only)	79.54	77.35	78.44(77.3 ± 17.2)	76.09	48.77	62.43(64.2 ± 21.5)
CBrain	79.63	78.01	**78.82**(78.6 ± 15.7)	74.84	52.27	63.56(63.2 ± 22.3)

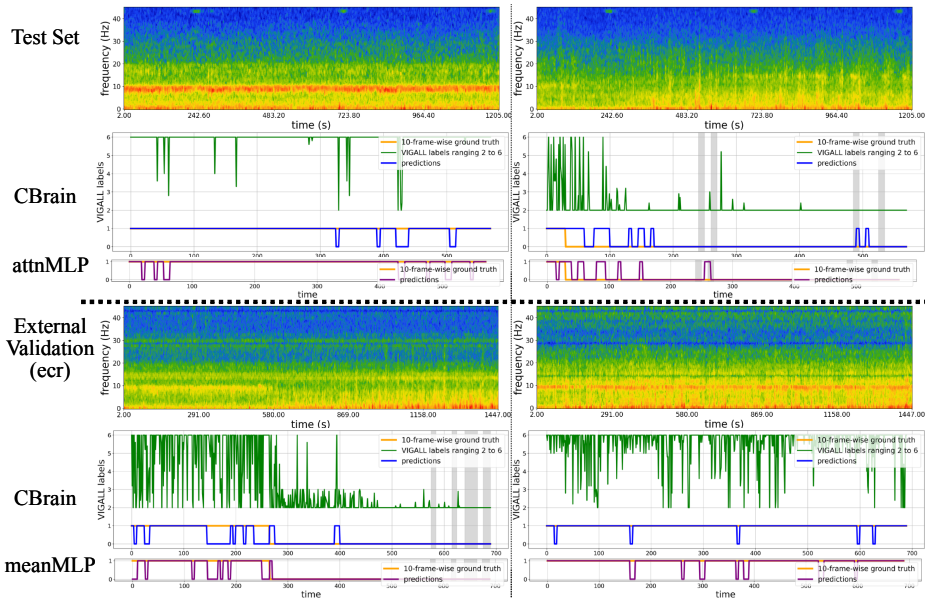

Fig. 2. Visualization of CBrain's predictions on the testing set and external validation (eye-closed-rest) scans, along with the best-performing baselines (testing set: attnMLP[20], ecr scans: meanMLP[20]). Top plots of each panel: EEG spectrograms[22]. Bottom plots of each panel: model predictions (blue), baseline predictions (purple), the 10-frame ground truth (orange), the original frame-wise integer Vigilance Algorithm Leipzig (VIGALL) labels (green), noisy EEG segments (gray bars). Higher labels indicate higher alertness levels[19]. The overlap between predicted and ground truth labels reflects strong performance.

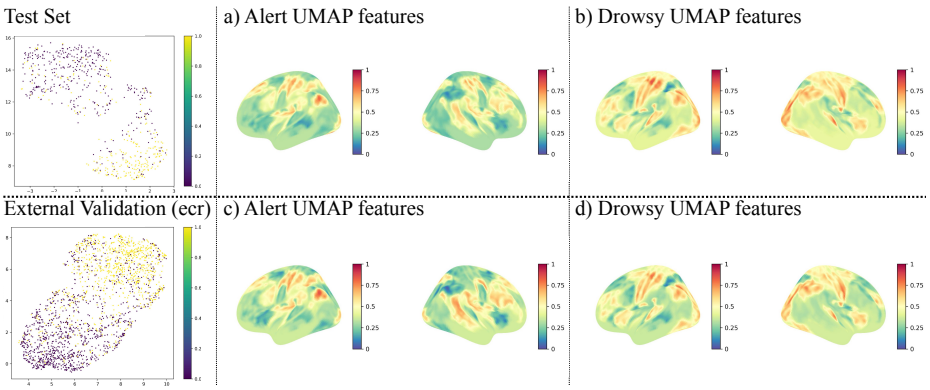

Fig. 3. Qualitative analysis of CBrain's extracted fMRI features. Each row illustrates the UMAP[16] embedding of fMRI features within the dataset, and the mapping of average alert and drowsy fMRI features on the brain surface in both hemispheres.

4 Conclusion

We propose **CBrain**: **C**ross-modal learning for **Brain** vigilance detection from resting-state fMRI, which transfers cross-modal vigilance knowledge from EEG to the fMRI domain. Our model accurately predicts vigilance state from fMRI data in a 10-fMRI-frame granularity. This approach can provide vigilance-state annotations to fMRI scans, including public databases, in the common scenario where dedicated vigilance measures (such as EEG and eye behavior) are not simultaneously recorded. Our model's robustness in generalization tasks supports the idea that cross-modal EEG information can enhance the discriminability of brain state patterns within single-modal fMRI data.

Acknowledgments. This work was supported by NIH grants R01 NS112252, F99 AG079810, and T32 EB021937. The external dataset was acquired in the Advanced MRI Section of the NINDS, NIH.

Disclosure of Interests. The authors have no competing interests.

References

1. Bedel, H.A., Sivgin, I., Dalmaz, O., Dar, S.U., Çukur, T.: Bolt: Fused window transformers for fmri time series analysis. Medical image analysis **88**, 102841 (2023)
2. Caro, J.O., Fonseca, A.H.d.O., Averill, C., Rizvi, S.A., Rosati, M., Cross, J.L., Mittal, P., Zappala, E., Levine, D., Dhodapkar, R.M., et al.: Brainlm: A foundation model for brain activity recordings. bioRxiv pp. 2023–09 (2023)
3. Chang, C., Leopold, D.A., Schölvinck, M.L., Mandelkow, H., Picchioni, D., Liu, X., Ye, F.Q., Turchi, J.N., Duyn, J.H.: Tracking brain arousal fluctuations with fmri. Proceedings of the National Academy of Sciences **113**(16), 4518–4523 (2016)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Dadi, K., Varoquaux, G., Machlouzarides-Shalit, A., Gorgolewski, K.J., Wassermann, D., Thirion, B., Mensch, A.: Fine-grain atlases of functional modes for fmri analysis. NeuroImage **221**, 117126 (2020)
6. Dong, Z., Li, R., Wu, Y., Nguyen, T.T., Chong, J.S.X., Ji, F., Tong, N.R.J., Chen, C.L.H., Zhou, J.H.: Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. arXiv preprint arXiv:2409.19407 (2024)
7. Falahpour, M., Chang, C., Wong, C.W., Liu, T.T.: Template-based prediction of vigilance fluctuations in resting-state fmri. Neuroimage **174**, 317–327 (2018)
8. Goodale, S.E., Ahmed, N., Zhao, C., de Zwart, J.A., Özbay, P.S., Picchioni, D., Duyn, J., Englot, D.J., Morgan, V.L., Chang, C.: fmri-based detection of alertness predicts behavioral response variability. elife **10**, e62376 (2021)
9. Jiang, W.B., Zhao, L.M., Lu, B.L.: Large brain model for learning generic representations with tremendous eeg data in bci. arXiv preprint arXiv:2405.18765 (2024)
10. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G.: Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage **146**, 1038–1049 (2017)

11. Kostas, D., Aroca-Ouellette, S., Rudzicz, F.: Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. Frontiers in Human Neuroscience **15**, 653659 (2021)
12. Li, Y., Lou, A., Xu, Z., Zhang, S., Wang, S., Englot, D.J., Kolouri, S., Moyer, D., Bayrak, R.G., Chang, C.: Neurobolt: Resting-state eeg-to-fmri synthesis with multi-dimensional feature mapping. arXiv preprint arXiv:2410.05341 (2024)
13. Liu, T.T., Falahpour, M.: Vigilance effects in resting-state fmri. Frontiers in neuroscience **14**, 321 (2020)
14. Liu, Y., Wu, H., Wang, J., Long, M.: Non-stationary transformers: Exploring the stationarity in time series forecasting. Advances in Neural Information Processing Systems **35**, 9881–9893 (2022)
15. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-vocabulary point-cloud object detection without 3d annotation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1190–1199 (2023)
16. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
17. Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., et al.: Multimodal population brain imaging in the uk biobank prospective epidemiological study. Nature neuroscience **19**(11), 1523–1536 (2016)
18. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2906–2917 (2021)
19. Olbrich, S., Fischer, M.M., Sander, C., Hegerl, U., Wirtz, H., Bosse-Henck, A.: Objective markers for sleep propensity: comparison between the multiple sleep latency test and the vigilance algorithm leipzig. Journal of sleep research **24**(4), 450–457 (2015)
20. Popov, P., Mahmood, U., Fu, Z., Yang, C., Calhoun, V., Plis, S.: A simple but tough-to-beat baseline for fmri time-series classification. NeuroImage **303**, 120909 (2024)
21. Pourmotabbed, H., Martin, C.G., Goodale, S.E., Doss, D.J., Wang, S., Bayrak, R.G., Kang, H., Morgan, V.L., Englot, D.J., Chang, C.: Multimodal state-dependent connectivity analysis of arousal and autonomic centers in the brainstem and basal forebrain. bioRxiv pp. 2024–11 (2024)
22. Prerau, M.J., Brown, R.E., Bianchi, M.T., Ellenbogen, J.M., Purdon, P.L.: Sleep neurophysiological dynamics through the lens of multitaper spectral analysis. Physiology **32**(1), 60–92 (2017)
23. Tagliazucchi, E., von Wegner, F., Morzelewski, A., Borisov, S., Jahnke, K., Laufs, H.: Automatic sleep staging using fmri functional connectivity data. Neuroimage **63**(1), 63–72 (2012)
24. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.M.H., et al.: The wu-minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
25. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
26. Wang, C., Ong, J.L., Patanaik, A., Zhou, J., Chee, M.W.: Spontaneous eyelid closures link vigilance fluctuation with fmri dynamic connectivity states. Proceedings of the National Academy of Sciences **113**(34), 9653–9658 (2016)

27. Wang, G., Liu, W., He, Y., Xu, C., Ma, L., Li, H.: Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
28. Wang, Y., Huang, N., Li, T., Yan, Y., Zhang, X.: Medformer: A multi-granularity patching transformer for medical time-series classification. arXiv preprint arXiv:2405.19363 (2024)
29. Wei, Z., Dan, T., Ding, J., Wu, G.: Neuropath: A neural pathway transformer for joining the dots of human connectomes. arXiv preprint arXiv:2409.17510 (2024)
30. Yang, C., Westover, M., Sun, J.: Biot: Biosignal transformer for cross-data learning in the wild. Advances in Neural Information Processing Systems **36** (2024)
31. Zhang, S., Goodale, S.E., Gold, B.P., Morgan, V.L., Englot, D.J., Chang, C.: Vigilance associates with the low-dimensional structure of fmri data. NeuroImage **267**, 119818 (2023)