A Foundation Model for Brain MRI with Dynamic Modality Integration (Student Abstract)

Minh Sao Khue Luu, Bair N. Tuchinov

The Artificial Intelligence Research Center of Novosibirsk State University
1 Pirogova Street, Novosibirsk, 630090, Russian Federation
khue.luu@g.nsu.ru, bairt@nsu.ru

Abstract

We introduce a single–backbone foundation model for brain MRI that supports dynamic modality integration: it operates with arbitrary, possibly unseen, combinations of MRI sequences at pretrain and transfer. The encoder is conditioned by text derived modality embeddings via conditional layer normalization, while a variance–covariance penalty discourages feature collapse. Unlike expert-based designs that grow with each new sequence, our approach scales without adding modality-specific branches. Pretrained self-supervised on ∼60,000 heterogeneous MRIs, the model learns modality-aware yet modality-agnostic features. We outline evaluation on segmentation and classification under missing/unseen modalities and cross-center shifts, and present early feasibility on multiple sclerosis lesion segmentation under limited data. This work moves toward robust, protocol-agnostic MRI foundation models suited to real clinical variability.

Code — https://github.com/BrainFM/brainfm

Extended version —
https://doi.org/10.48550/arXiv.2511.03014

Introduction

Foundation models have transformed NLP, computer vision, and multimodal learning, yet their extension to medical imaging remains nontrivial. Brain MRI depends on multiple complementary sequences (e.g., T1, T2, FLAIR), but acquisition protocols vary widely across scanners and centers. This variability causes inconsistent modality availability and limits the generalization of architectures trained on fixed modality sets (Zhang et al. 2022). Quantitative analyses of public brain MRI datasets further reveal large differences in modality composition, underscoring the need for models that can flexibly adapt to heterogeneous and incomplete MRI inputs (Luu et al. 2025).

Several recent efforts have explored this direction. AMAES (Munk et al. 2024) pretrains 3D backbones via masked reconstruction on ∼45k brain MRIs in single-modality settings. M4oE (Jiang and Shen 2024) and MoME (Zhang et al. 2024) use modality-specific experts with gating for fusion; new MRI sequences require adding an expert and fine-tuning. mmFormer (Zhang et al. 2022) handles missing modalities in brain tumor segmentation via modality-specific and cross-modal Transformers with auxiliary regularization. These approaches either assume fixed modality sets or expand architectural complexity as modalities increase, limiting their scalability and practicality.

We present BrainFM-MRI, a modality-conditioned masked autoencoder for large-scale 3D brain MRI pretraining. The model integrates three key components: (1) Modality-Conditioned Patch Encoding, which combines text-derived modality embeddings with conditional layer normalization for modality-aware feature adaptation; (2) Padding-Aware Masking, ensuring stable reconstruction under missing or irregular modality inputs; and (3) Variance–Covariance Regularization, promoting representation diversity and preventing feature collapse. BrainFM-MRI aims to provide a single, flexible backbone that remains stable across different MRI protocols, taking a step toward modality-agnostic foundation models.

Methodology

Modality-Conditioned Patch Encoding Each MRI volume is divided into non-overlapping 3D patches, which are flattened and linearly projected into token embeddings. Each token is augmented with (i) a text-derived modality embedding from the frozen BioBERT encoder (Lee et al. 2020), projected to the token dimension, and (ii) a learnable 3D positional embedding.

To inject modality context throughout the network, we replace LayerNorm with Conditional Layer Normalization (CLN) (Chen et al. 2021). Given token features $\mathbf{h}_{s,i}$ and a modality embedding $\mathbf{e}_m$, CLN applies modality-dependent scaling and shifting:

$$\mathrm{CLN}(\mathbf{h}_{s,i} \mid \mathbf{e}_m) = \gamma(\mathbf{e}_m) \cdot \frac{\mathbf{h}_{s,i} - \mu}{\sigma} + \beta(\mathbf{e}_m). \quad (1)$$

Because patch and modality embeddings share the same token space, and CLN modulates features without altering the architecture, the encoder learns representations that remain stable across different and even unseen modality combinations. A high-level schematic is shown in Figure 1.

Padding-Aware Masking Different inputs may contain different numbers of MRI sequences. After patchifying each modality and concatenating all tokens, sessions with fewer

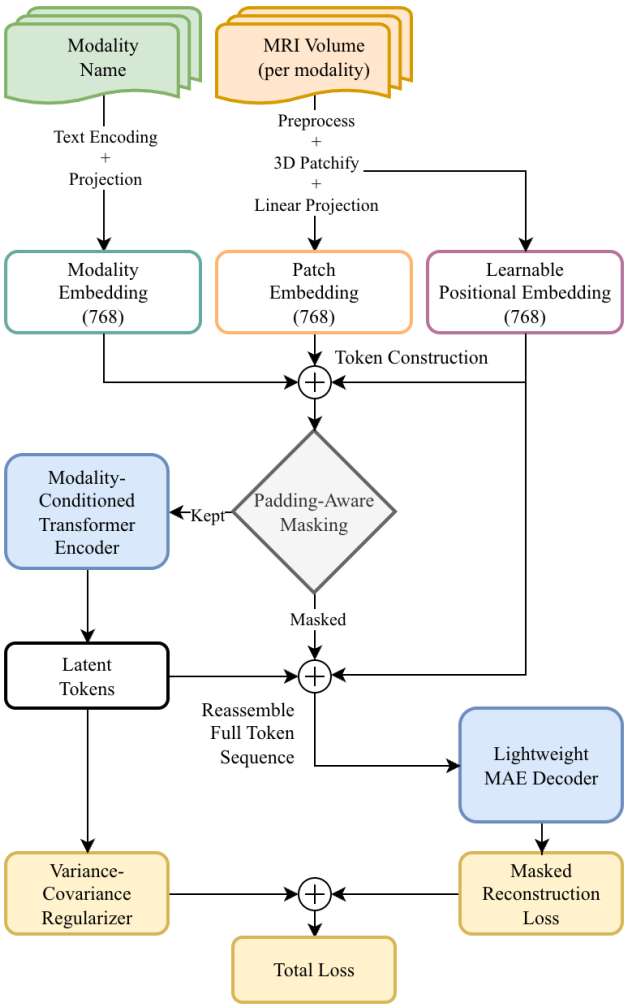

Figure 1: Overview of the proposed modality-conditioned masked autoencoder for 3D brain MRI. Visible (unmasked) patches are encoded, and masked patches are reconstructed in the decoder. The workflow includes patch extraction, modality conditioning, masking, and reconstruction.

modalities are padded to match the batch sequence length. These padded positions are ignored by the encoder, decoder, and loss. Masking is sampled only from valid (non-padded) positions, preventing missing modalities from being selected and keeping a stable number of visible tokens. This ensures the encoder always processes clean anatomical tokens and supports representation learning that is consistent across varying modality sets. A lightweight Transformer decoder, following the standard MAE design (He et al. 2021), takes the visible encoder tokens and modality-conditioned mask tokens as input, and reconstructs the masked patches back to their original spatial locations.

Variance–Covariance Regularization To reduce feature collapse, we apply a VICReg-inspired penalty (Bardes, Ponce, and LeCun 2021). Encoder features $\mathbf{z} \in \mathbb{R}^{N \times D}$ with $N = B \times L_{\text{keep}}$ (total retained tokens across the batch) are constrained by two auxiliary losses: (i) a *variance* term encouraging per-dimension standard deviation ≥ 1, and (ii) a *covariance* term penalizing redundant cross-dimension correlations:

$$\mathcal{L}_{\text{var}} = \frac{1}{D} \sum_{j} \text{ReLU}\left(1 - \sqrt{\text{Var}(z_{\cdot j}) + \epsilon}\right), \quad (2)$$

$$\mathcal{L}_{\text{cov}} = \frac{1}{D} \sum_{i \neq j} \text{Cov}(z_{\cdot i}, z_{\cdot j})^2. \quad (3)$$

Pretraining Objective The final objective combines masked reconstruction and representation regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MAE}} + \lambda_{\text{var}} \mathcal{L}_{\text{var}} + \lambda_{\text{cov}} \mathcal{L}_{\text{cov}}, \quad (4)$$

with $\lambda_{\text{var}} = 0.1$ and $\lambda_{\text{cov}} = 0.003$ linearly warmed up over the first five epochs. Reconstruction is computed using MSE over masked, valid voxel elements only.

Experimental Plan

Benchmark We compare BrainFM-MRI against AMAES, MoME, M4oE, and mmFormer, using identical preprocessing, augmentations, and compute budgets for fairness. For the preliminary stage, we only test feasibility against nnU-Net (Isensee et al. 2021) to verify the framework under limited data.

Evaluation Tasks Segmentation: lesion and tumor segmentation with robustness to missing or unseen modalities and cross-domain variation. Classification: subject-level or scan-level prediction tasks using (i) frozen-encoder linear probes, (ii) shallow head fine-tuning, and (iii) full-model fine-tuning where applicable. Metrics include Dice, HD95, sensitivity, and specificity.

Data We use the FOMO25 dataset (Munk et al. 2025) for pretraining. It includes 11,187 subjects, 13,900 sessions, and 60,529 MRI scans in total. Each subject may undergo one or more imaging sessions, and each session contains several MRIs saved in the NIfTI format. We treat an input at the session level.

Preliminary Results

To verify the feasibility of our framework, we conducted a small-scale experiment on 10 subjects from the MSLesSeg dataset (Guarnera et al. 2025). Training was limited to two epochs using only T1 and FLAIR modalities. Table 1 reports early Dice scores and Hausdorff distances, comparing BrainFM-MRI with nnU-Net. Even under this short training regime, our model shows consistent improvements when modalities are missing.

Model	**Dice** ↑	**HD95** ↓
nnU-Net (T1+FLAIR)	0.39	12.8
BrainFM-MRI (ours)	0.45	11.2

Table 1: Preliminary results on 10 MSLesSeg cases. Full-scale experiments are ongoing.

Acknowledgments

This work was supported by a grant for research centers, provided by the Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement with the Novosibirsk State University dated April 17, 2025 No. 139-15-2025-006: IGK 000000C313925P3S0002.

References

Bardes, A.; Ponce, J.; and LeCun, Y. 2021. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. Version Number: 3.

Chen, M.; Tan, X.; Li, B.; Liu, Y.; Qin, T.; Zhao, S.; and Liu, T.-Y. 2021. AdaSpeech: Adaptive Text to Speech for Custom Voice. Version Number: 1.

Guarnera, F.; Rondinella, A.; Crispino, E.; Russo, G.; Di Lorenzo, C.; Maimone, D.; Pappalardo, F.; and Battiato, S. 2025. MSLesSeg: baseline and benchmarking of a new Multiple Sclerosis Lesion Segmentation dataset. *Scientific Data*, 12(1): 920.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked Autoencoders Are Scalable Vision Learners. Version Number: 3.

Isensee, F.; Jaeger, P. F.; Kohl, S. A. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2): 203–211.

Jiang, Y.; and Shen, Y. 2024. M4oE: A Foundation Model for Medical Multimodal Image Segmentation with Mixture of Experts. In Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; and Schnabel, J. A., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15012, 621–631. Cham: Springer Nature Switzerland. ISBN 978-3-031-72389-6 978-3-031-72390-2. Series Title: Lecture Notes in Computer Science.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Luu, M. S. K.; Benedichuk, M. V.; Roppert, E. I.; Kenzhin, R. M.; and Tuchinov, B. N. 2025. A Structured Review and Quantitative Profiling of Public Brain MRI Datasets for Foundation Model Development. Version Number: 1.

Munk, A.; Ambsdorf, J.; Llambias, S.; and Nielsen, M. 2024. AMAES: Augmented Masked Autoencoder Pretraining on Public Brain MRI Data for 3D-Native Segmentation. Version Number: 2.

Munk, A.; Cerri, S.; Ambsdorf, J.; Machnio, J.; Llambias, S. N.; Nersesjan, V.; Krag, C. H.; Liu, P.; García, P. R.; Ghazi, M. M.; Boesen, M.; Benros, M. E.; Iglesias, J. E.; and Nielsen, M. 2025. A large-scale heterogeneous 3D magnetic resonance brain imaging dataset for self-supervised learning. ArXiv:2506.14432 [eess].

Zhang, X.; Ou, N.; Basaran, B. D.; Visentin, M.; Qiao, M.; Gu, R.; Ouyang, C.; Liu, Y.; Matthews, P. M.; Ye, C.; and Bai, W. 2024. A Foundation Model for Brain Lesion Segmentation with Mixture of Modality Experts. In Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; and Schnabel, J. A., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15012, 379–389. Cham: Springer Nature Switzerland. ISBN 978-3-031-72389-6 978-3-031-72390-2. Series Title: Lecture Notes in Computer Science.

Zhang, Y.; He, N.; Yang, J.; Li, Y.; Wei, D.; Huang, Y.; Zhang, Y.; He, Z.; and Zheng, Y. 2022. mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation. In Wang, L.; Dou, Q.; Fletcher, P. T.; Speidel, S.; and Li, S., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13435, 107–117. Cham: Springer Nature Switzerland. ISBN 978-3-031-16442-2 978-3-031-16443-9. Series Title: Lecture Notes in Computer Science.