A Foundation Model for Brain MRI with Dynamic Modality Integration

Minh Sao Khue Luu[1] and Bair N. Tuchinov[1]

[1]The Artificial Intelligence Research Center of Novosibirsk State University, Novosibirsk 630090, Russia
khue.luu@g.nsu.ru

Abstract

We present a foundation model for brain MRI that can work with different combinations of imaging sequences. The model uses one encoder with learnable modality embeddings, conditional layer normalization, and a masked autoencoding objective that accounts for missing modalities. A variance–covariance regularizer is applied to stabilize feature learning and improve representation diversity. This design removes the need for separate models for each modality and allows the network to adapt when some sequences are missing or unseen. It is trained on about 60,000 multi-center MRIs using self-supervised reconstruction and modality imputation to learn flexible representations. A learnable modality embedding guides feature extraction so the encoder can adjust to different inputs. We describe our planned evaluation on brain tumor and multiple sclerosis segmentation, as well as lesion classification, under various modality settings. Preliminary results show that the method works feasibly, and further experiments are planned to study its performance in more detail. All code and pretrained models are available at https://github.com/BrainFM/brainfm

Keywords: foundation models; self-supervised learning; MRI; segmentation

1 Introduction

Foundation models (FMs) have transformed natural language processing, computer vision, and multimodal learning, yet their application to medical imaging remains challenging. Brain MRI relies on multiple sequences (e.g., T1, T2, FLAIR), but acquisition protocols vary widely across sites. This variation leads to inconsistent modality availability and limits the generalization of fixed-modality architectures [1].

Masked autoencoders and foundation models have demonstrated strong capabilities for learning general-purpose visual representations [2]. In medical imaging, self-supervised pretraining approaches such as Models Genesis [3] have shown that large-scale unsupervised learning can reduce annotation costs and improve cross-dataset generalization. Recent work extends these ideas to multimodal MRI: AMAES [4] pretrains on 44k MRIs using masked autoencoding but supports only single-modality inputs; M4oE [5] and MoME [6] use modality-specific experts with gating to enable dynamic fusion but require new experts for additional sequences. mmFormer [1] addresses missing MRI modalities for brain tumor segmentation using modality-specific and cross-modal Transformers with regularization for robust predictions.

Existing approaches such as AMAES are limited to single-modality inputs, while MoME and M4oE rely on separate expert networks for each sequence. Our model uses a single backbone combined with learnable modality embeddings, allowing it to handle new sequences without adding new experts. Compared with mmFormer, which focuses on task-specific robustness,

we pretrain a unified encoder with modality imputation for broader downstream applications. The proposed foundation model for brain MRI (a) conditions feature extraction on learnable text-derived modality embeddings, (b) adapts shared encoder parameters through conditional layer normalization, and (c) integrates self-supervised modality imputation within a masked reconstruction framework.The design removes the need for modality-specific branches and maintains robustness when sequences are missing or previously unseen. A variance–covariance regularizer further stabilizes learning across variable modality sets, enabling consistent performance across diverse MRI configurations.

2 Method

2.1 Dataset Preparation and Preprocessing

Pretraining Dataset. We use the FOMO25 dataset [7] 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. In this study, we treat a *case* at the session level: for example, if a subject has two sessions, these are considered as two separate cases.

Training Dictionary. To manage the data efficiently, we organize all subject sessions in a nested dictionary (subjects with multiple sessions have distinct keys). The outer dictionary uses `case_id` as the key, and its value is another dictionary that maps each modality to the corresponding NIfTI image, i.e., `{case_id : {modality : path}}`. This structure allows flexible handling of cases with different numbers or types of modalities. This nested structure allows flexible access to heterogeneous modality combinations across sessions.

Preprocessing Pipeline. During data loading, all modalities of a given session are read from disk and cropped or padded to a consistent spatial size of $(128, 128, 128)$ (voxels). A sequence of 3D data augmentations is then applied using the `MONAI` library. Table 1 summarizes all preprocessing operations and their hyperparameters. To improve efficiency, text embeddings of modality names are cached using a modality encoder cache, ensuring that identical modality tokens share the same embedding across samples.

2.2 Architecture

BrainFM is a modality-conditioned masked autoencoder designed for large-scale 3D brain MRI pretraining. Its design unifies four key ideas: (1) learnable text-derived modality embeddings that generalize to unseen sequences, (2) conditional layer normalization (CLN) for adaptive feature modulation across modalities, (3) padding-aware masking for handling variable input completeness, and (4) variance–covariance regularization for stable and diverse representation learning.

Overview. Each MRI volume is divided into non-overlapping 3D patches that are flattened and linearly projected into tokens. Each token is enriched with a modality embedding (obtained from a pretrained text encoder) and a learnable 3D positional embedding. The modality embedding is injected at each encoder layer through CLN, allowing modality-specific feature scaling and shifting during representation learning. This conditioning allows the encoder to adaptively modulate its normalization statistics depending on the modality. A padding-aware masking strategy (randomly hides a fixed proportion ρ of valid patches while ensuring that padded or empty regions are never selected for masking) is then applied. The visible tokens are processed by a Transformer encoder with CLN, and the masked tokens are reconstructed by a lightweight decoder using modality-aware mask tokens and positional re-injection, as shown in Figure 1.

Table 1: Preprocessing and augmentation operations applied during dataset loading.

Operation	**Description and Hyperparameters**
`DivisiblePad`	Pads volumes to make each spatial dimension divisible by the patch size $(16, 16, 16)$.
`RandBiasField`	Random bias field distortion; probability $p = 0.3$, coefficient range $(0.3, 0.6)$.
`RandGaussianNoise`	Adds Gaussian noise; probability $p = 0.3$, mean 0.0, standard deviation 0.05.
`RandAdjustContrast`	Random contrast adjustment; probability $p = 0.3$, gamma range $(0.7, 1.5)$.
`RandFlip`	Random flipping along all spatial axes; probability $p = 0.5$.
`RandAffine`	Random rotation within $\pm 15^\circ$ about each axis ($\pm\pi/12$ radians); padding mode: `border`.
`NormalizeIntensity`	Normalizes nonzero voxels per channel to zero mean and unit variance.
`Sanitize Voxels`	Replaces non-finite values (NaN, Inf) with zeros and clamps range to $[-4, 4]$.

Masked Reconstruction Loss. The model is first trained to reconstruct the hidden (masked) patches from the visible ones. The objective is a mean squared error (MSE) loss computed only on masked and valid voxel elements:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N} \sum_{s} \sum_{i \in \mathcal{H}_s} \sum_{p} \left(\widehat{P}_{s,i}^{(p)} - P_{s,i}^{(p)} \right)^2, \tag{2.1}$$

where $P_{s,i}^{(p)}$ and $\widehat{P}_{s,i}^{(p)}$ denote original and reconstructed voxel intensities, $\mathcal{H}_s$ are masked patch indices, and N is the number of valid voxel elements. This reconstruction task encourages the model to learn modality-aware and spatially coherent features across heterogeneous MRI inputs.

Variance–Covariance Regularization. To further stabilize pretraining and avoid feature collapse, BrainFM-MRI applies a VICReg-inspired regularization [8]. Each sample produces a mean-pooled feature vector $\mathbf{z}_b \in \mathbb{R}^D$, stacked across the batch as $\mathbf{z} \in \mathbb{R}^{B \times D}$. Two auxiliary losses are added: the *variance term*, which enforces unit variance per feature dimension, and the *covariance term*, which penalizes correlations between features:

$$\mathcal{L}_{\text{var}} = \frac{1}{D} \sum_{j} \text{ReLU}\left(1 - \sqrt{\text{Var}(z_{\cdot j}) + \epsilon}\right) \tag{2.2}$$

$$\mathcal{L}_{\text{cov}} = \frac{1}{D(D-1)} \sum_{i \neq j} \text{Cov}(z_{\cdot i}, z_{\cdot j})^2. \tag{2.3}$$

Total Objective. The final pretraining objective combines reconstruction and regularization terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MAE}} + \lambda_{\text{var}} \mathcal{L}_{\text{var}} + \lambda_{\text{cov}} \mathcal{L}_{\text{cov}}, \tag{2.4}$$

where $\lambda_{\text{var}} = 0.1$ and $\lambda_{\text{cov}} = 0.005$ are linearly increased during the first five epochs (warm-up).

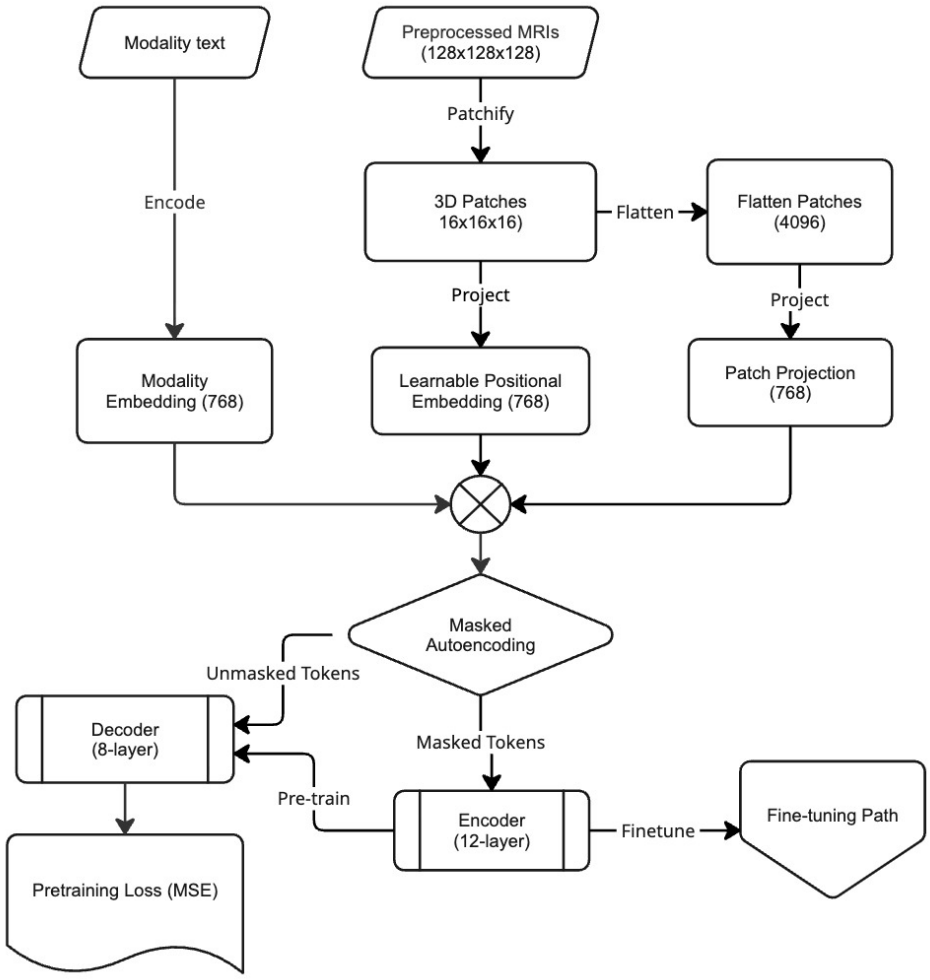

Figure 1: Overview of the proposed modality-conditioned masked autoencoder for 3D brain MRI. Visible (unmasked) patches are encoded, and masked patches are reconstructed in the decoder. The workflow includes patch extraction, modality conditioning, masking, and reconstruction.

Transfer to Downstream Tasks. After pretraining, the decoder is discarded, and the encoder is adapted to downstream segmentation or classification tasks. A lightweight head is attached—mean-pooled for classification or upsampled for segmentation— and trained with a standard supervised loss $\mathcal{L}_{\text{task}}$.

3 Experimental Plan

Models. We compare BrainFM against AMAES, MoME, M4oE, and mmFormer. All models use identical preprocessing, augmentations, and compute budgets for fairness.

Datasets. We evaluate on two segmentation datasets (e.g., BraTS-MET, MSLesSeg) and two classification datasets (e.g., lesion presence/severity or tumor subtype) plus one cross-domain dataset to probe shift. Each dataset is split subject-wise into train/val/test (70/10/20) with three fixed, non-overlapping folds; no subject leakage. All images are resampled to a common voxel spacing, intensity normalized (z-score), and optionally bias-field corrected. Spatial (flip, affine) and photometric (noise, contrast) augmentations are matched across models. Patch size, stride, and masking (for MAE-style methods) are harmonized wherever applicable.

Table 2: Example modality configurations used for evaluation. X indicates that the modality is available, and – indicates it is missing.

Configuration	T1	T1c	T2	FLAIR
Complete	X	X	X	X
Dropped (T1c)	X	–	X	X
Dropped (T2)	X	X	–	X
Dropped (FLAIR)	X	X	X	–
Unseen (T1+FLAIR only)	X	–	–	X
Unseen (T2 only)	–	–	X	–

Training Protocol. All models use AdamW, cosine decay with linear warm-up, the same batch size, epochs, and weight decay. Early stopping is based on validation Dice (segmentation) or AUROC (classification). Each run repeats with three seeds; we report mean ± std. Hyperparameters are tuned on the first fold's validation split, then fixed for all other folds.

Evaluation Tasks. Segmentation: lesion/tumor segmentation with robustness to missing/unseen modalities and cross-domain generalization. Classification: subject-level or scan-level labels (e.g., lesion presence, tumor subtype, disease vs. healthy), with: (i) frozen-encoder linear probe, (ii) shallow head finetuning, and (iii) full-model finetuning (where permitted).

Modality-Availability Protocol. All models are evaluated under the configurations in Table 2. "Unseen" rows are held out from training. For BrainFM-MRI, we additionally perform zero-shot tests on unseen modality sets to assess compositional generalization.

4 Preliminary Results

This small-scale experiment is an initial step toward the full benchmark described in Section 3. In this experiment, we pretrain the model using masked autoencoding with modality-conditioned reconstruction. For transfer, we consider two tasks: brain tumor segmentation and multiple sclerosis lesion analysis. Robustness is assessed using a Modality-Availability Matrix that includes full, missing, and unseen sequence configurations. Baselines include nnU-Net, SegResNet, and ablations without modality embeddings or imputation. Metrics follow standard practice (Dice, Hausdorff distance, sensitivity, specificity). Experiments are ongoing, and complete results will be reported in the final version.

At this stage, we verify the encoder's feasibility on a small 10-case subset of the MSLesSeg dataset. Training was limited to two epochs using only the T1 and FLAIR modalities. Table 3 shows early Dice and Hausdorff distance results, comparing the modality-conditioned MAE (BrainFM) with nnU-Net. These results illustrate early evidence that the modality-conditioned encoder learns stable representations even under limited supervision.

Table 3: Preliminary results on 10 MSLesSeg cases.

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

Table 4: Comparison of related foundation models.

Model	Input	Handling	Scalability
AMAES	Single	Masked AE	Limited
MoME	Multi	Experts + Gating	Needs new experts
M4oE	Multi	Mixture of Experts	Scales poorly
mmFormer	Multi	Cross-modal Transformer	Robust
Ours	Arbitrary	Embeddings + CLN	Scalable

5 Conclusion and Discussion

We presented BrainFM-MRI, a modality-conditioned foundation model for brain MRI that unifies masked autoencoding and modality embeddings within a single encoder. Unlike prior expert-based architectures, BrainFM-MRI adapts dynamically to any set of MRI sequences, including missing or unseen ones. Table 4 compares our approach with representative foundation models for brain MRI. Preliminary results indicate the feasibility of the approach, though experiments are still ongoing. Future work will extend training to the full dataset and include comprehensive evaluations across multiple tasks.

Acknowledgements. 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

[1] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13435, pages 107–117. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-16442-2 978-3-031-16443-9. doi: 10.1007/978-3-031-16443-9_11. URL https://link.springer.com/10.1007/978-3-031-16443-9_11. Series Title: Lecture Notes in Computer Science.

[2] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, 2021. URL https://arxiv.org/abs/2111.06377. Version Number: 3.

[3] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. Models Genesis. 2020. doi: 10.48550/ARXIV.2004.07882. URL https://arxiv.org/abs/2004.07882. Publisher: arXiv Version Number: 4.

[4] Asbjørn Munk, Jakob Ambsdorf, Sebastian Llambias, and Mads Nielsen. AMAES: Augmented Masked Autoencoder Pretraining on Public Brain MRI Data for 3D-Native Segmentation, 2024. URL https://arxiv.org/abs/2408.00640. Version Number: 2.

[5] Yufeng Jiang and Yiqing Shen. M4oE: A Foundation Model for Medical Multimodal Image Segmentation with Mixture of Experts. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15012, pages

621–631. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-72389-6 978-3-031-72390-2. doi: 10.1007/978-3-031-72390-2_58. URL https://link.springer.com/10.1007/978-3-031-72390-2_58. Series Title: Lecture Notes in Computer Science.

[6] Xinru Zhang, Ni Ou, Berke Doga Basaran, Marco Visentin, Mengyun Qiao, Renyang Gu, Cheng Ouyang, Yaou Liu, Paul M. Matthews, Chuyang Ye, and Wenjia Bai. A Foundation Model for Brain Lesion Segmentation with Mixture of Modality Experts. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15012, pages 379–389. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-72389-6 978-3-031-72390-2. doi: 10.1007/978-3-031-72390-2_36. URL https://link.springer.com/10.1007/978-3-031-72390-2_36. Series Title: Lecture Notes in Computer Science.

[7] Asbjørn Munk, Stefano Cerri, Jakob Ambsdorf, Julia Machnio, Sebastian Nørgaard Llambias, Vardan Nersesjan, Christian Hedeager Krag, Peirong Liu, Pablo Rocamora García, Mostafa Mehdipour Ghazi, Mikael Boesen, Michael Eriksen Benros, Juan Eugenio Iglesias, and Mads Nielsen. A large-scale heterogeneous 3D magnetic resonance brain imaging dataset for self-supervised learning, June 2025. URL http://arxiv.org/abs/2506.14432. arXiv:2506.14432 [eess].

[8] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning, 2021. URL https://arxiv.org/abs/2105.04906. Version Number: 3.