SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Evaluating Vision-Language Foundation Models for Brain MRI Sequence Classification, Stability, and Metadata Harmonization

Satvik Tripathi, Perelman School of Medicine at University of Pennsylvania; Arnold Campbell, PhD

Introduction/Background

To comprehensively evaluate the diagnostic performance, inference stability, and scalability of vision-language foundation models (VLMs) in zero-shot classification of brain MRI sequences. This study explores their utility for automated sequence labeling and metadata harmonization in both curated test sets and large-scale clinical datasets.

Methods/Intervention

We evaluated five state-of-the-art VLMs—BiomedGPT, BiomedCLIP, LLaVA-Med, MedCLIP, and UniMedCLIP—on 2D slices from brain MRI. An initial benchmark test set included 25 images spanning six canonical sequences: T1-weighted, T1 post-contrast, T2-weighted, T2-FLAIR, diffusion-weighted imaging (DWI), and susceptibility-weighted imaging (SWI). For stability testing, each model underwent 10 repeated zero-shot inference trials per image, and consistency was quantified using statistical analysis of modal prediction frequency. To assess real-world generalizability, UniMedCLIP was further evaluated on (1) a curated 140-image dataset across seven classes and (2) a substantially larger multi-site, multi-scanner dataset of 571 images spanning 13 diverse MRI sequence types. Performance was assessed via overall accuracy, per-class accuracy, and confusion matrices.

Results/Outcome

On the benchmark set, UniMedCLIP achieved the highest accuracy (90.2%) and demonstrated robust sequence-wise performance. Stability testing revealed significant differences between models (ANOVA $p < 0.001$, $\eta^2 = 0.760$), with UniMedCLIP showing the highest consistency (mean = 9.76/10, SD = 0.44), achieving perfect predictions in 76% of cases. On the 140-sequence test set, UniMedCLIP reached 92.1% accuracy. On the 571-sequence set, it achieved 82% overall accuracy, with high performance across common classes and modest confusion between visually similar sequences (e.g., DWI vs. ADC, FLAIR variants).

Conclusion

UniMedCLIP consistently demonstrates high zero-shot classification accuracy, reproducibility, and scalability across both curated and real-world datasets. These zero-shot results support the use of domain-specialized VLMs for automated MRI sequence labeling and metadata harmonization, though task-based fine-tuning may help achieve sufficient clinical accuracy.

Statement of Impact

VLMs offer a scalable and reliable solution for harmonizing MRI metadata, reducing manual labeling burden, and improving dataset interoperability—key steps toward robust AI deployment in radiology.

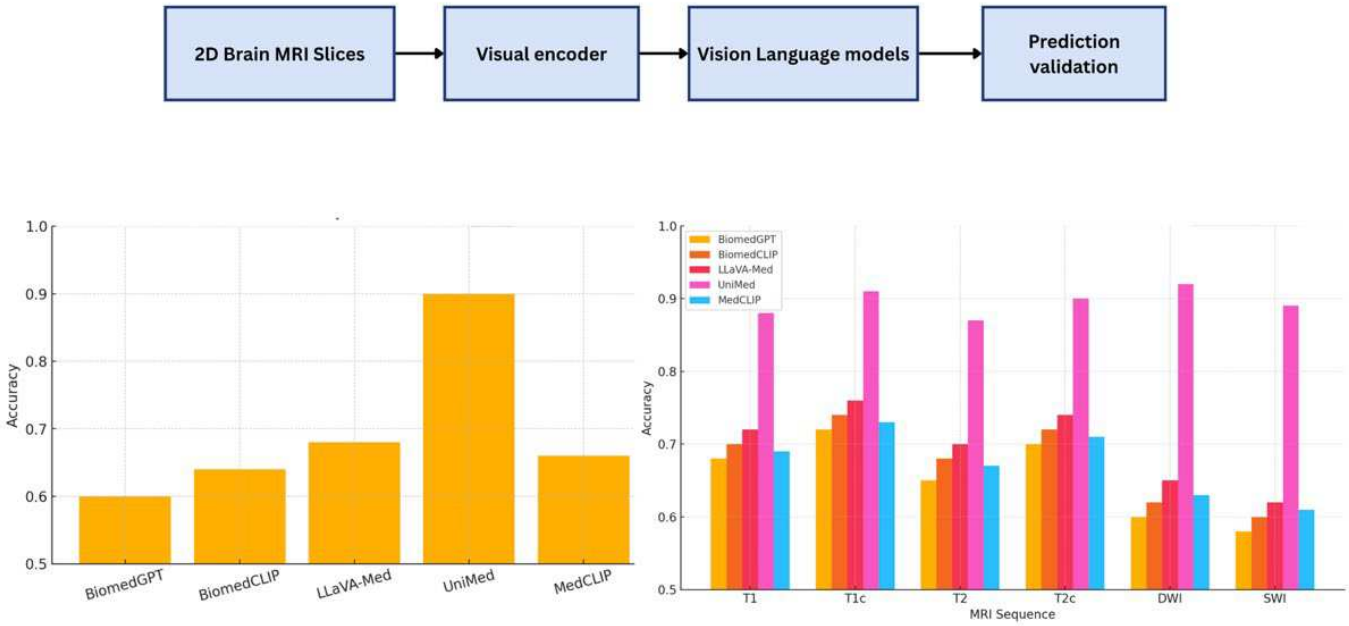
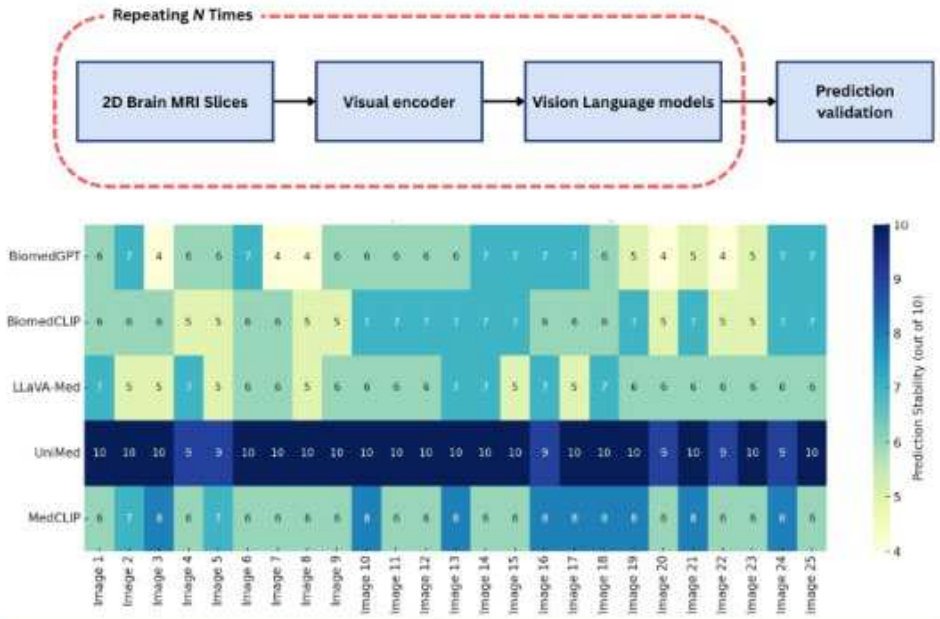


Figure.1 Evaluation of Vision-Language Models (VLMs) for Zero-Shot Classification of Brain MRI Sequences. (Top) Schematic of the VLM pipeline: single-slice 2D brain MRI inputs are passed through a visual encoder and processed by vision-language models to classify MRI sequence type, followed by accuracy-based validation. (Bottom Left) Overall classification accuracy of five VLMs—BiomedGPT, BiomedCLIP, LLaVA-Med, UniMed, and MedCLIP—demonstrates highest performance by UniMed. (Bottom Right) Per-sequence accuracy across six MRI types (T1, T1c, T2, T2 Flair, DWI, SWI), showing that UniMed consistently outperforms other models across all sequences. Notably, BiomedGPT and MedCLIP show comparatively lower and more variable performance.



Model	Mean Accuracy (%)	Std Dev (%)	Min Accuracy (%)	Max Accuracy (%)
BiomedGPT	58.0	11.2	40.0	70.0
BiomedCLIP	61.2	8.3	50.0	70.0
LLaVA-Med	60.0	7.1	50.0	70.0
UniMed	97.6	4.4	90.0	100.0
MedCLIP	68.0	9.6	60.0	80.0

Figure 2. Stochastic Stability Evaluation of Vision-Language Models (VLMs) for Brain MRI Sequence Classification. (Top) Workflow schematic showing repeated inference pipeline using 2D brain MRI slices with five vision-language models to assess stability across 25 images. (Middle) Heatmap displaying the number of consistent predictions (out of 10 runs) per image for each model. UniMed demonstrates near-perfect consistency, while other models show greater variability across runs. (Bottom) Summary table of prediction stability metrics across models. UniMed achieved the highest mean accuracy (97.6%) with the lowest standard deviation (4.4%), highlighting its robustness. In contrast, BiomedGPT showed the lowest mean accuracy (58.0%) and highest variability (SD = 11.2%).

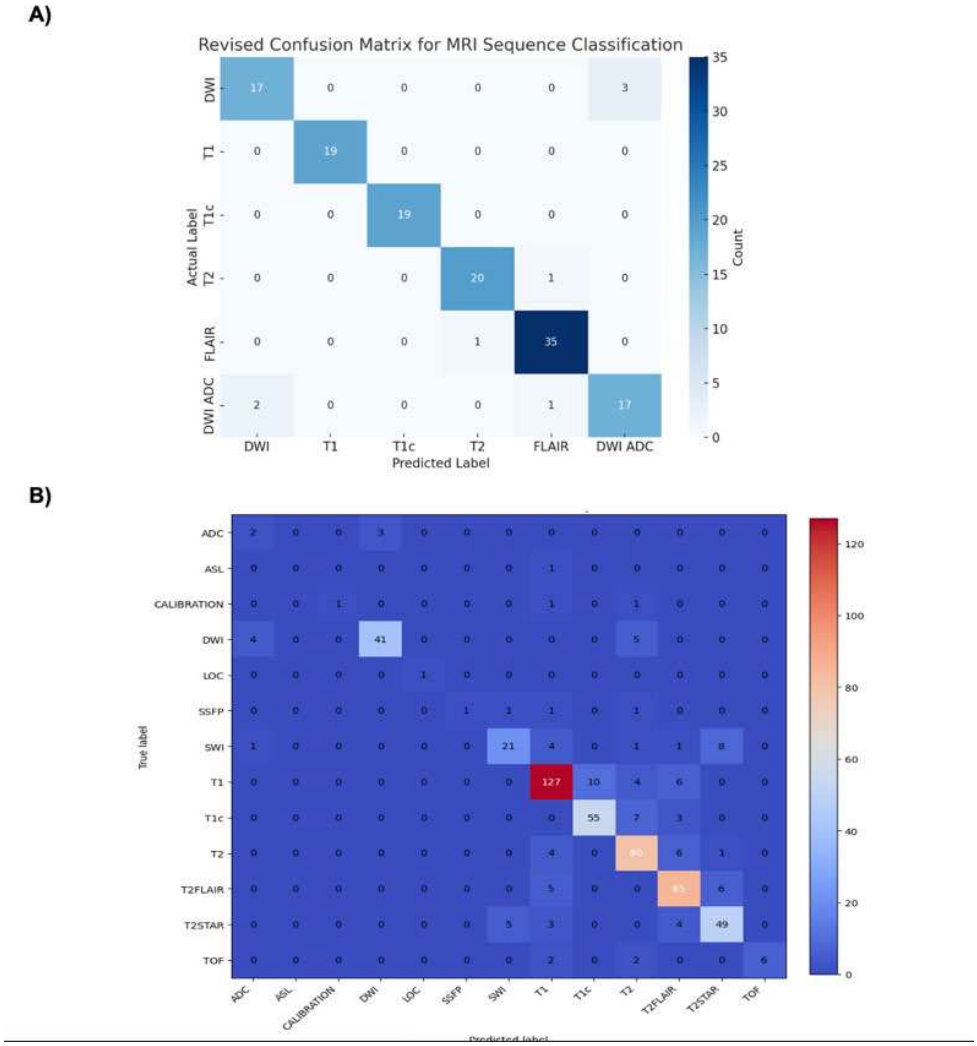


Figure.3. Zero-Shot Classification Performance of UniMedCLIP on Curated and Large-Scale Brain MRI Datasets. (A) Confusion matrix for predictions on a curated dataset of 140 brain MRI sequences spanning seven common sequence types. Most misclassifications occurred between semantically or visually similar sequences such as DWI vs. DWI ADC and FLAIR variants. (B) Confusion matrix on an expanded institutional dataset of 571 sequences covering 13 distinct MRI sequence types. UniMedCLIP demonstrates strong performance across frequent classes (e.g., T1, T2, T2-FLAIR), with modest confusion among related or low-frequency sequences. These results highlight the model's robustness and generalizability in harmonizing diverse MRI metadata without fine-tuning.

Keywords

Vision-Language Model; Brain MRI; Metadata Harmonization**;** AI