A Vision-Language Foundation Model for Precise and Comprehensive Brain Tumor Diagnosis from Preoperative Multimodal Data

Yinong Wang [†,a], Jianwen Chen [†a], Zhou Chen, MD [†b], Shuwen Kuang, MD [†c], Haoning Jiang [a], Yanzhao Shi [v], Huichun Yuan, MD [d], Yan-ran (Joyce) Wang, PhD [e], Bing Wang, MD [g], Lei Wu, MD [h], Bin Tang, MD [i], Li Meng, MD [j], Baihua Luo, MD [k], Bin Zhou, MD [f,l], Wei Ding, MD [m], Weiming Zhong, MD [n], Wei Hou, MD [o], Yuanbing Chen, MD [p], Zhiping Wan, MD [q], Wei Wang, MD [r], Zhenkun Xiao, MD [g], Wenwu Wan, MD [s], Allen He[t], Yuyin Zhou, PhD [u], Prof Longbo Zhang, MD [b,f,x], Feifei Wang, PhD [v], Prof Zhixiong Liu, MD [b,f,x], Prof Michael Iv, MD [w], Xuan Gong, MD PhD [b,f,x,*], Liangqiong Qu, PhD [a,*]

[a] *School of Computing and Data Science, The University of Hong Kong, China*
[b] *Department of Neurosurgery, Xiangya Hospital, Central South University, China*
[c] *Department of Oncology, Xiangya Hospital, Central South University, China*
[d] *Changde Hospital, Xiangya School of Medicine, Central South University (The First People's Hospital of Changde City), China*
[e] *Department of Biomedical Data Science, School of Medicine, Stanford University, Stanford, USA*
[f] *Department of Neurosurgery, Xiangya Hospital, Central South University, Jiangxi (National Regional Center for Neurological Diseases), China*
[g] *Department of Neurosurgery, The Second Affiliated Hospital, Hengyang Medical School, University of South China, China*
[h] *Department of Neurosurgery, The Second Affiliated Hospital, Jiangxi Medical College, Nanchang University, China*
[i] *Department of Neurosurgery, The First Affiliated Hospital, Jiangxi Medical College, Nanchang University, China*
[j] *Department of Radiology, Xiangya Hospital, Central South University, China*
[k] *Department of Pathology, Xiangya Hospital, Central South University, China*
[l] *Department of Neurosurgery, Jiangxi Provincial People's Hospital, The First Affiliated Hospital of Nanchang Medical College, China*
[m] *The Affiliated Children's Hospital of Xiangya School of Medicine, Hunan Children's Hospital, China*
[n] *Department of Neurosurgery, Shenzhen Second People's Hospital, China*
[o] *Department of Neurosurgery, First Hospital of Lanzhou University, China*
[p] *Department of Neurosurgery, The Third Xiangya Hospital, Central South University, China*
[q] *Department of Neurosurgery, Tongji Hospital, School of Medicine, Tongji University, China*
[r] *Department of Radiology, Tongji Hospital, School of Medicine, Tongji University, China*
[s] *Department of Neurosurgery, Chongqing Traditional Chinese Medicine Hospital, China*
[t] *Basis International School Park Lane Harbour, China*
[u] *Department of Computer Science and Engineering, University of California, Santa Cruz, USA*
[v] *Department of Electrical and Electronic Engineering, The University of Hong Kong, China*
[w] *Department of Radiology, School of Medicine, Stanford University, Stanford, USA*
[x] *National Clinical Research Center of Geriatric Disorders, Xiangya Hospital, Central South University, China*

* Corresponding authors
Email addresses: gong.xuan@csu.edu.cn (Xuan Gong), liangqqu@hku.hk (Liangqiong Qu)
† These authors contributed equally to this work.

Summary

Background Non-invasive presurgical diagnosis of brain tumor types from Magnetic Resonance Imaging (MRI) is essential but challenging due to overlapping imaging features across tumor types, inter-observer variability, and the extensive training required for expertise. We aimed to develop an MRI-based Artificial Intelligence (AI) model for automatic and reliable brain tumor classification with diagnostic uncertainty quantification and radiology reports generation.

Methods We developed BrainVLM to classify all 12 World Health Organization (WHO) 2021 brain tumor types. BrainVLM integrates an uncertainty quantification strategy to indicate prediction reliability and a module for generating radiology reports to elucidate the clinical rationale. BrainVLM was trained on multi-modal data (MRI scans, demographics, and radiology reports) from 40,043 individuals. It was validated on 5,211 patients with pathologically confirmed brain tumors, including 3,877 held-out patients from the primary hospital and 1,334 patients from 11 independent hospitals. We further conducted two proof-of-concept studies to validate its clinical utility in AI-clinician workflows: 1) a blinded multi-reader study where 12 neuroradiologists across varying experience levels interpreted 248 retrospective cases with or without AI assistance, and 2) a real-world prospective study in which 1,009 patients were independently and blindly assessed by BrainVLM and radiologists before surgery. Additionally, we demonstrated BrainVLM's utility in preoperative molecular subgroup prediction for adult-type diffuse gliomas, using a multi-center cohort of 632 patients.

Findings In primary evaluation, BrainVLM achieved an area under the curve (macro-AUC) of 0.85 (95% CI: 0.84–0.86), and an F1 score of 0.82 (95% CI: 0.81–0.83), surpassing neuroradiologists (F1 = 0.80 (95% CI: 0.79–0.81)). In external validation across 11 centers, BrainVLM achieved an AUC = 0.80 (95% CI: 0.79–0.82) and F1 = 0.75 (95% CI: 0.73–0.78), compared with F1=0.71 (95% CI: 0.69-0.73) for neuroradiologists. In prospective real-world evaluation, BrainVLM maintained performance comparable to neuroradiologists. In the blinded multi-reader study, the AI-assisted neuroradiologists demonstrated a statistically 27.6% (absolute 0.16, 95% CI: 20%-35%) improvement in F1 score across all experience levels compared to unaided assessments ($p<0.0001$), while reducing diagnostic time by 34.7% (absolute 55.8 seconds, 95% CI: 29.9%-38.9%). Across 5,211 cases, 72% (3,008/4,156) of correct diagnoses had high-confidence scores (>90%), whereas 70% (736/1,055) of incorrect diagnoses exhibited low-confidence scores (50–85%). For molecular subgroup prediction in adult-type diffuse gliomas, BrainVLM achieved AUCs of 0.95 (95% CI: 0.93-0.96) and 0.88 (95% CI: 0.84-0.91) in primary and external validation, respectively.

Interpretations By integrating uncertainty quantification and report generation, BrainVLM improves radiologists' diagnostic performance, reduces MRI interpretation time, and allows lower-confidence cases to be effectively flagged for human review. These findings show that AI-assisted diagnosis based on presurgical MRI can support clinical decision-making, with the potential to enhance early brain tumor diagnosis, alleviate clinicians' workloads, and ultimately improve patient care.

Research in context

Evidence before this study

We searched PubMed and Google Scholar for English-language studies published up to December 6, 2025, using the terms "MRI" AND ("brain tumor" OR "CNS tumor" OR "brain cancer") AND ("machine learning" OR "artificial intelligence" OR "deep learning" OR "foundation model") in titles or abstracts. We identified no reports describing the deployment of an MRI-based AI model capable of diagnosing the full spectrum of the 12 major brain tumor types defined by WHO CNS5 (the 2021 fifth edition of the WHO Classification of Tumors of the Central Nervous System). Existing literature on AI-assisted brain tumor diagnosis focused almost exclusively on gliomas, meningiomas, and brain metastases, with little attention to other WHO CNS tumor types. Critically, to our knowledge, no previous work has rigorously quantified model reliability or provided clinically interpretable diagnostic rationales (such as radiology style reports) alongside brain tumor classifications. As a result, existing AI outputs for brain tumor classification are limited to isolated predictions, lacking the clinical narrative components that are imperative for clinician trust and real-world adoption.

Added value of this study

To our knowledge, this is the first study to develop and validate an AI model capable of comprehensive classification across the full spectrum of all 12 major brain tumor types from preoperative multimodal data. By leveraging the largest and most diverse multi-modal neuro-oncology dataset to date, our model achieved robust diagnostic performance with high generalizability, superior or comparable to board-certified neuroradiologists in sensitivity, precision, and F1 scores on both internal and multi-center external validation cohorts. By integrating diagnostic precision with automatic radiology reports generation and reliable uncertainty quantification, our BrainVLM could substantially improve neuroradiologists' diagnostic accuracy and efficiency across a wide spectrum of tumor types, supporting neuroradiologists at all levels of experience. We publicly release our source code and our training dataset to promote transparency, reproducibility, and future research in the field of neuro-oncology AI.

Implications of all the available evidence

Use of AI foundation models, when trained on sufficiently large and diverse datasets, has the potential to substantially improve the accuracy and efficiency of preoperative tumor diagnosis across a broad spectrum of tumor types, benefiting radiologists at all levels of experience. Reliable uncertainty quantification and automated report generation enhance the interpretability and clinical utility of AI predictions, helping to build clinician trust and streamline decision-making. Prospective clinical studies are warranted to confirm the real-world benefits and safety of integrating AI-assisted diagnostics into standard neuro-oncological management.

Introduction

Brain neoplasms can have long-lasting and life-altering physical, cognitive, and psychological impacts on patients' lives, and can occur in any anatomical region of the encephalon and adjacent structures.[1–3] WHO CNS5 identifies 12 major brain tumor types and over 100 distinct

tumor subtypes, ranging from benign and indolent neoplasms such as grade 1 meningioma to malignant and aggressive tumors such as grade 4 glioblastoma.[4] Common treatment options for brain tumors include neurosurgical resection, radiotherapy, and chemotherapy, with surgery serving as the primary initial management in most cases. However, the choice of treatment modality, surgical approach, and the extent of resection depends largely on the initial presurgical diagnosis.[5] For example, maximal safe gross total resection is generally recommended for patients with diffuse gliomas. In contrast, lymphomas and some germ cell tumors may be managed effectively with chemotherapy and/or radiotherapy without surgical resection.[6,7] Hence, a prompt and accurate preoperative diagnosis of brain tumors is an important first step in guiding appropriate treatment.

MRI is the primary diagnostic modality for presurgical diagnosis and longitudinal monitoring of patients with brain tumors.[8] In routine practice, neuroradiologists must review multi-parametric MRI scans to identify and diagnose potential lesions and then convey their findings and impression in a radiology report. However, this manual interpretation is time-consuming and error-prone due to overlapping radiologic characteristics of different brain tumors, inter-observer variability, and potential perceptual errors. This problem is further compounded by the global shortage of specialist neuroradiologists due to the extensive training efforts required to gain expertise. Together, these unmet needs call for advanced techniques to develop an automatic brain tumor diagnostic tool for rapid and precise diagnoses and report generation, thereby accelerating the time-sensitive diagnosis process and reducing clinician workload.

AI, particularly the recently emerged vision language models (VLMs),[9] has the potential to advance automatic disease diagnosis. Yet, compared to the flourishing development of AI diagnostic frameworks for other tumors, comprehensive AI models for presurgical brain tumor diagnosis remain significantly underexplored. At least two critical challenges hinder the widespread clinical adoption of AI-based brain tumor diagnosis. First, current publicly available datasets for brain tumor analysis are limited in scope, predominantly featuring gliomas, meningiomas, and metastases, while underrepresenting other tumors in the WHO CNS5.[10] Second, existing AI models predict brain tumor types without assessing the confidence and reliability of the prediction or explaining the underlying clinical rationale. This issue is particularly pronounced in large language models (LLMs) or VLMs, which often exhibit unwarranted high confidence in erroneous predictions.[11] This overconfidence in AI predictions, coupled with the opaque decision-making process, undermines clinicians' trust and poses a substantial barrier to their clinical translation and adoption.

In this study, we developed BrainVLM, an MRI-based AI model for precise and automated classification of all 12 WHO CNS5-defined brain tumor categories. For each case, BrainVLM outputs a diagnosis, a confidence score to quantify diagnostic reliability, and a radiology report to explain the clinical rationale. We compared BrainVLM's performance against board-certified neuroradiologists and state-of-the-art AI models, using pathological diagnoses and radiology reports as reference standards. Furthermore, we validated its clinical utility in two proof-of-concept AI–clinician workflow studies: 1) a blinded multi-reader study in which 12 neuroradiologists of varying experience levels interpreted 248 retrospective cases with and without AI assistance, and 2) a real-world prospective study in which 1009 patients were assessed simultaneously by BrainVLM and radiologists prior to surgery. Finally, we

demonstrate BrainVLM's utility in preoperative molecular subgroup prediction for adult-type diffuse gliomas using multi-center cohorts.

Methods

Datasets and study design

For the development of BrainVLM, we curated BrainTumor48K, a comprehensive multi-modal brain tumor dataset. It comprises 47,947 individuals aggregated from 39 public repositories and 12 collaborating medical centers (Figure 1a, Table 1, and Supplementary Tables 1-4, appendix pp 46-50). The BrainTumor48K comprises 33,149 patients with pathologically confirmed brain tumors (21,993 from public repositories and 11,156 from medical centers) and 14,798 healthy controls (sourced from public repositories) (Figure 1a). The detailed public repositories and their pre-processing pipeline for the public repositories are summarized in the Supplementary Section S.1.1.1 (appendix pp 2-3). Data from 12 independent collaborating medical centers included demographics, 3D multi-parametric MRI scans (T1-weighted (T1), T1 contrast-enhanced (T1c), T2-weighted (T2), and T2-Flair sequences (T2f)), expert-curated radiology reports, and pathological diagnoses.

Overall, the BrainTumor48K dataset spans all 12 major brain tumor types and most of the subtypes defined by the 2021 WHO CNS5.[4] The detailed number of patients across the 12 major brain tumor types and their subtypes is shown in the right panel of Figure 1a.

Data from participating medical institutions were obtained with ethics approval from their respective Institutional Review Boards (IRBs), and the details of these IRBs are presented in the Supplementary Section S1.3 (appendix pp 6-7). The requirement for informed consent was waived by the IRBs. All data were de-identified in compliance with institutional policies and ethical standards. This study complies with the TRIPOD+AI guidelines for the reporting artificial intelligence-based prediction models.[12]

Development of AI foundation model and retrospective validation

BrainVLM (Figure 1b) was trained on a dataset aggregating 35,227 publicly available cases (including patients with brain tumors and healthy controls) and 4,816 patients from the primary institutional cohort (Xiangya Hospital). A validation set, consisting of 500 public cases and 120 primary cohort cases, was reserved for hyperparameter tuning and model selection. Retrospective diagnostic performance was evaluated on 5,211 patients with pathologically confirmed brain tumors, including 3,877 held-out patients from the primary hospital, and 1,334 patients from 11 independent hospitals (Figure 1a). More details regarding the BrainVLM network architecture, uncertainty quantification, training strategies and validation procedures can be found in the Supplementary Sections S2 (appendix pp 8-25).

Prospective validation procedures

For the prospective study, we consecutively enrolled 1162 patients with an initial diagnosis of brain lesions at Xiangya Hospital and additional 349 consecutive patients from two external institutional cohorts (Supplementary Tables 3, 4, appendix pp 49-50). Following brain MRI acquisition, clinical information and MRI images were collected and preprocessed according to standardized protocols. Patients were excluded if they had a history of prior brain surgery,

previous brain radiotherapy or stereotactic radiosurgery, missing MRI images, or incomplete MRI protocols. Following the application of these criteria (Figure 2c), eligible patients were divided into two groups: surgical group (n = 886; 639 Xiangya, 247 external) and non-operative group (n = 123; 74 Xiangya, 49 external). Eligible patients' data were prospectively predicted using BrainVLM prior to definitive clinical diagnosis. Following patient discharge, BrainVLM outputs were validated against gold standard references: postoperative pathological findings for patients who underwent surgery, or consensus discharge diagnoses established by a multidisciplinary panel of senior clinicians through comprehensive review of radiological and clinical data for non-surgical cases. Finally, patients with non-tumoral pathological findings were also excluded.

Multi-reader study for AI-augmented clinical assessments

Multi-reader studies are widely adopted to assess the clinical validity of medical AI models.[13] To achieve a more rigorous and realistic evaluation of BrainVLM's clinical utility in augmenting neuroradiological diagnosis, we developed a gold-standard test dataset comprising 248 patients covering 12 major WHO CNS5 brain tumor types. For each case, readers received only the multi-parametric MRI scans (T1, T1c, T2, T2-Flair) and patient metadata (age, sex); no clinical history was provided. Diagnostic performance was systematically compared across three paradigms: (a) neuroradiologists-only as a reader, (b) BrainVLM as an autonomous diagnostic agent, and (c) neuroradiologists augmented by BrainVLM as a reader. In paradigm (c), BrainVLM assisted neuroradiologists by providing diagnoses with associated confidence scores and a radiology report. Twelve neuroradiologists, stratified by expertise (5 juniors: 3-5 years; 4 seniors: 5-10 years; 3 experts: 10+ years), were recruited in the blinded crossover study. For each case interpretation, neuroradiologists recorded a diagnosis and a self-rated diagnostic confidence score. To minimize memorization effects in scenarios (a) and (c), we reshuffled the order of the cases and modified their anonymized identifiers for the second read, and allowed a one-month wash-out period between the two reads.

Statistical analysis

BrainVLM was compared against board-certified neuroradiologists (using clinical radiology reports) and four AI models, with postoperative pathology as the gold standard. RadFM,[14] Merlin,[15] and VST[16] were fine-tuned on BrainTumor48K, while ChatGPT-4o[17] served as a zero-shot generalist baseline. Tumor classification performance was quantified using sensitivity, specificity, precision, Cohen's Kappa, F1, and the area under the curve (AUC). For multi-label classification, we computed the class-frequency-weighted F1 and both macro- and micro-averaged AUC. Report generation was assessed with BLEU-4[18], RaTEScore[19], RadGraph-XL[20], and an LLM-as-Judge framework (Qwen 2.5-72B). We applied two-tailed Wilcoxon signed-rank tests to assess statistical significance and reported 95% confidence intervals. Following previous work, we applied non-parametric bootstrap resampling (1,000 replicates) to estimate 95% CIs.[21] P values were not adjusted for multiple comparisons. The performance metrics in the benchmark analyses (Figures 1d and 3d) and tumor categories in the multi-reader study (Figures 4a-d) were interpreted as distinct analytical objectives, rather than repeated tests of a single hypothesis. Therefore, no Bonferroni or other multiplicity correction was applied. We performed Shapley value analysis on the primary and external test datasets to evaluate

BrainVLM's robustness to missing MRI sequences (Supplementary Section S3.2, appendix pp 25-26).

Role of funding source

The funders of this study had no role in data collection, analysis, interpretation, writing of the manuscript, and the decision to submit the manuscript for publication.

Results

Diagnostic performance of BrainVLM in primary and external test datasets

We first evaluated BrainVLM's diagnostic ability using the retrospective primary test dataset (n = 3,877). BrainVLM achieved a macro-AUC of 0.85 (95% CI: 0.84–0.86, Figure 1c; 0.87 for both intra-axial and extra-axial tumors, Supplementary Figure 10b, appendix p 65) and a weighted-average F1 score of 0.82 (95% CI: 0.81–0.83), with precision of 0.84 (95% CI: 0.83–0.85) and sensitivity of 0.82 (95% CI: 0.81–0.83; Figure 1d). This exceeded expert neuroradiologists' assessments (F1 = 0.80, 95% CI: 0.79–0.81; precision = 0.71, 95% CI: 0.70–0.73; sensitivity = 0.80, 95% CI: 0.79–0.81), with enhanced inter-rater agreement (Cohen's κ = 0.75, 95% CI: 0.74–0.76 vs 0.69, 95% CI: 0.68–0.70). In benchmarking against four comparator AI models (RadFM, Merlin, VST, and ChatGPT-4o), BrainVLM improved key performance metrics by at least 25%, including in several low-prevalence tumor types (Figure 1d; Supplementary Table 11, appendix p 56). DeLong tests showed that BrainVLM had a higher AUC than each comparison model (Supplementary Table 5 and Supplementary Figure 10, appendix pp 51, 65).

In the external multicentric dataset, BrainVLM maintained robust performance. It achieved a macro-AUC of 0.80 (95% CI: 0.79–0.82, Figure 1c; 0.76 for intra-axial and 0.84 for extra-axial tumors, Supplementary Figure 10b, appendix p 65) and an F1 score of 0.75 (95% CI: 0.73–0.78; precision = 0.77, 95% CI: 0.75–0.78; sensitivity = 0.75, 95% CI: 0.74–0.77; Figure 1d). These results also surpassed both neuroradiologist consensus (F1 = 0.71, precision = 0.72, sensitivity = 0.71) and state-of-the-art models (RadFM: F1 = 0.43, Merlin: F1 = 0.52, VST: F1 = 0.41, ChatGPT-4o: F1 = 0.30).

BrainVLM exhibited slight performance variation across different tumor types (Figure 2a). In the primary test dataset, BrainVLM matched or surpassed expert-level neuroradiologists' performance for common brain tumors: MEN (F1 = 0.91 vs 0.91), GGN (F1 = 0.82 vs 0.81), and CPN (F1 = 0.78 vs 0.74). For low-prevalence or imaging-ambiguous brain tumors, BrainVLM showed superiority: HEM (F1 = 0.64 vs 0.45), EMB (F1 = 0.68 vs 0.58), and CPT (F1 = 0.55 vs 0.47). Confusion matrix analysis demonstrated substantial concordance between BrainVLM and human experts in diagnostic errors, with over 90% overlap in the top two misclassified tumor categories across both the primary test and external datasets (Figure 2b).

Notably, BrainVLM achieved robust performance using only standard MRI sequences (T1, T1c, T2, T2-Flair) and demographic data, whereas neuroradiologists relied on supplemental advanced imaging (e.g., diffusion-weighted imaging, MR angiography, MR spectroscopy, and MR Perfusion) in 56.1% of cases (Supplementary Table 7, appendix p 53). Together, these results establish BrainVLM as an innovative AI system that achieves neuroradiologist-level performance on most brain tumor types across heterogeneous clinical settings without relying on advanced imaging or protocol harmonization.

Real-world prospective study

To further evaluate the utility of BrainVLM in real-world clinical settings, we conducted a multi-center prospective study aimed at assessing its performance in authentic diagnostic workflows. In the surgical group, BrainVLM achieved F1 scores of 0.78 (primary) and 0.80 (external), surpassing the radiologists' performance of 0.75 and 0.77, respectively (Figure 3a). In the non-operative group, the model maintained robust diagnostic performance, yielding F1 scores of 0.85 in the primary cohort and 0.75 in the external cohort (Figure 3a), both exceeding the neuroradiologists' benchmarks (0.79 and 0.72, respectively). These results demonstrate that BrainVLM provides reliable diagnostic support across diverse clinical management pathways and independent patient populations.

Uncertainty-aware clinical decision support

Conventional VLMs can generate incorrect diagnoses with inappropriately high confidence or fail to provide uncertainty quantification, potentially undermining clinicians' trust.[22] To address this issue, we developed a consensus-driven strategy to augment BrainVLM with reliable uncertainty quantification (Supplementary Section S2.2, appendix pp 16-18). Across 5,211 cases in the primary and external test datasets, 72% of correct diagnoses (n = 3,008/4,156) were associated with high-confidence (>90%), whereas 70% of incorrect diagnoses (n = 736/1,055) exhibited confidence scores of 50-85%. By contrast, the corresponding model without the consensus-driven strategy assigned low confidence to only 15% of incorrect predictions (n = 160/1,055; Figure 3c). Formal calibration analysis showed good agreement between predicted confidence and observed outcomes, with an expected calibration error (ECE)[23] of 0.036 and a Brier score[24] of 0.1448 (Figure 3b). This was further supported by the reliability diagram, in which observed diagnostic accuracy increased with predicted confidence (Figure 3b and Figure 3c).

We next examined whether these uncertainty estimates could inform decision support in diagnostically ambiguous cases. Cases assigned lower confidence by BrainVLM were associated with lower inter-reader agreement among neuroradiologists (Supplementary Section S3.8.2, appendix p 31), supporting the clinical relevance of the model's uncertainty estimates. We therefore implemented a confidence-triggered supplementary diagnosis mechanism in BrainVLM (Supplementary Section S2.3.1, appendix p 21): when the confidence of the top-ranked diagnosis (Top 1) was below 75%, BrainVLM additionally presented the second-ranked diagnosis (Top 2, Supplementary Figure 9a and d, appendix p 64) for radiologist review. Under this strategy, overall F1 improved from 0.82 (95% CI: 0.81–0.83) to 0.86 (95% CI: 0.85–0.87) in the primary dataset and from 0.75 (95% CI: 0.73–0.78) to 0.78 (95% CI: 0.75–0.80) in the external dataset (Supplementary Section 3.1, appendix p 25).

Radiology report generation

We next evaluated BrainVLM's ability to generate radiology reports to accompany its diagnostic predictions. We compared BrainVLM-generated reports with those produced by other VLMs. BrainVLM consistently outperformed all baselines across both primary and external test datasets (Figure 3d). In the primary dataset, BrainVLM achieved superior BLEU-4 (0.43 vs 0.02–0.38), F1RadGraph-XL (0.57 vs 0.15–0.46), and RaTEScore (0.75 vs 0.47–

0.63). Similarly, in the external dataset, BrainVLM maintained performance, with BLEU-4 (0.35 vs 0.02–0.29), F1RadGraph-XL (0.52 vs 0.22–0.42), and RaTEScore (0.69 vs 0.42–0.58). We further compared BrainVLM with other VLMs on two key components of MRI reporting: signal intensity and tumor location. BrainVLM achieved report accuracies of 0.86 for contrast enhancement, 0.80 for T1 signal intensity, and 0.81 for T2 signal intensity, with a lesion localization accuracy of 0.72; these values were higher than those of comparator VLMs (Figure 3f, Supplementary Section S3.4, Supplementary Figure 7d, appendix pp 27, 63).

Preoperative molecular subgroup prediction in adult-type diffuse gliomas

Adult-type diffuse gliomas constitute the predominant malignant tumors affecting the central nervous system,[25] comprising three main subtypes: IDH-mutant astrocytoma, IDH-mutant oligodendroglioma, and IDH-wildtype glioblastoma.[4] Accurate distinction among these subtypes is essential for optimizing treatment strategies and predicting patient outcomes.[26] We finetuned BrainVLM on available adult-type diffuse glioma cases from BrainTumor48K, with threefold cross-validation on the primary dataset (n = 494) and external validation on an independent multi-center cohort (n = 138, subtype distributions in Supplementary Section 2.4, appendix p 21). In primary dataset, BrainVLM achieved a macro-averaged AUC of 0.95 (95% CI: 0.93-0.96; Figure 3e), surpassing Merlin (AUC 0.85, 95% CI: 0.82-0.86) and RadFM (AUC 0.76, 95% CI: 0.74-0.78). In the external cohort, BrainVLM maintained robust performance (AUC 0.88, 95% CI: 0.84-0.91), substantially exceeding Merlin (AUC 0.73, 95% CI: 0.69-0.77) and RadFM (AUC 0.63, 95% CI: 0.58-0.67). These results underscore the exceptional performance of BrainVLM for molecular subtyping, highlighting its potential as a versatile foundation model for brain tumor classification.

AI-augmented clinical assessments

In the 248-case multi-reader study, BrainVLM, when evaluated independently, outperformed junior and senior neuroradiologists across brain tumor types and performed comparably with expert neuroradiologists (Figure 4a-d and Table 2). With BrainVLM assistance, neuroradiologists' performance improved across all seniority groups, with a 27.6% increase in mean F1 score and a 34.7% reduction in diagnostic time compared with unaided assessment (both $p<0.0001$). F1 scores increased from 0.52 to 0.67 in junior neuroradiologists, from 0.55 to 0.74 in senior neuroradiologists, and from 0.79 to 0.85 in expert neuroradiologists (Table 2), bringing junior and senior readers closer to unaided expert-level performance.

To characterize clinician–AI interaction in diagnostically challenging settings, we analyzed cases in the lowest quartile of mean reader-reported confidence during unaided assessment (Supplementary Section S3.10, appendix pp 34-35). In this 25% low-confidence subset, the most common diagnostic trajectory across all reader seniority groups was trajectory F (initial clinician and BrainVLM both correct, final diagnosis remains correct; Figure 4f). Trajectory B (initially incorrect diagnosis corrected by a correct AI suggestion) was also frequent, 15.2% / 27.5% / 29.2% of junior / senior / expert assessments. By contrast, trajectory C (initially correct diagnosis reversed by an incorrect AI suggestion) was the least frequent outcome across groups (1.3%, 1.0%, and 0%, respectively; Figure 4f).

To further assess reader responses to incorrect AI outputs, we restricted analysis to cases in which BrainVLM generated an incorrect diagnosis (n = 52 of 248; Supplementary Section

S3.10, appendix pp 34-35). In this subset, overall accuracy was similar before and after AI assistance (50.0% vs 50.9%), with performance maintained or slightly improved in senior and expert readers and a modest decrease in juniors (Figure 4g). This slightly improved performance maybe because many incorrect BrainVLM predictions were associated with low reliability scores (Figure 4e), which may have prompted readers to apply more caution when reviewing these outputs. Overall, these findings do not suggest systematic over-reliance on incorrect AI outputs, although junior neuroradiologists appeared more susceptible to adverse AI influence.

Discussion

In this study, we developed and clinically evaluated BrainVLM, a multimodal vision-language foundation model for presurgical classification of brain tumors across all 12 WHO CNS5-defined tumor types. Three findings are particularly noteworthy. First, BrainVLM showed strong and generalizable diagnostic performance in retrospective evaluation, achieving AUCs of 0.85 in the primary validation set and 0.80 in the external validation set, with sensitivity, precision, and F1 scores broadly comparable to those of board-certified neuroradiologists across most tumor types. In subgroup analysis (appendix pp 31-33), performance was broadly consistent across sex, MRI vendor, ethnicity, and most age subgroups. Second, in prospective real-world evaluation, BrainVLM achieved a statistically significant 4% higher F1 score than neuroradiologists, suggesting potential clinical utility beyond retrospective benchmarking. Third, the model extended image-based tumor classification towards molecular stratification, achieving a superior AUC on external validation (0.88) relative to state-of-the-art VLMs (0.63–0.73). Taken together, these findings suggest that multimodal foundation models could support diagnostically demanding tasks in neuro-oncology, where overlapping MRI phenotypes, inter-observer variability, and limited familiarity with rare entities continue to constrain consistent presurgical diagnosis.[27]

The robust diagnostic performance of BrainVLM is likely attributable to the scale and diversity of its training dataset, and to its multimodal design. BrainTumor48K represents, to our knowledge, one of the largest and most comprehensive multi-modal presurgical datasets utilized for model development in neuro-oncology to date. Pretraining on extensive corpora comprising public repositories and real-world clinical data might have enabled BrainVLM to accommodate variability in imaging quality, acquisition parameters, patient populations, imaging equipment, and institutional protocols. Another key factor contributing to BrainVLM's performance is its efficient integration of multi-modal data, including MRI scans, patients' demographics, and radiological reports. This framework has enabled us to leverage critical clinical information such as age and radiology reports, leading to more accurate and comprehensive diagnoses.

AI-clinician collaboration holds significant promise in medical image interpretation.[28,29] However, even when AI models exhibit robust diagnostic performance and generalizability, translating these capabilities into clinical practice necessitates overcoming a critical barrier: clinician trust in AI-driven decision-making. The effective integration of AI models into clinical workflows requires clinicians to (1) understand the AI's decision rationale to inform their judgments, and (2) develop sufficient trust in the system's recommendations. To address these

challenges, BrainVLM introduces two clinician-trust-focused innovations: a consensus-driven uncertainty quantification strategy and an automated radiology report generation module. In our evaluation, BrainVLM demonstrated clinically meaningful self-assessment capabilities, assigning high-confidence probabilities (>90%) to 72% of its correct diagnoses while appropriately tagging 70% of its incorrect diagnoses with lower confidence scores 50-85%. This self-awareness capability is particularly impactful in diagnostically challenging scenarios. This mechanism ensures only high-confidence predictions are clinically actionable, while automatically triaging ambiguous cases to human experts for secondary review. This approach also prevents clinicians from over-reliance on model outputs, thereby aligning with responsible clinical workflows in healthcare settings. Complementing this reliability mechanism, BrainVLM exhibits an exceptional capability to generate preliminary medical reports—a critical function for communicating diagnostic findings and guiding care across several specialties.[28,30] These reports offer clinicians actionable insights beyond isolated predictions, enabling clinicians to process cases more efficiently, thereby improving turnaround times and expanding access to specialty-level reporting. The clinical validation through a multi-reader study with 12 neuroradiologists across expertise levels further confirmed BrainVLM's trust-building potential.

In clinical practice, incomplete MRI acquisition (e.g., missing T1c) frequently occurs due to insufficient clinician awareness or contraindications. While conventional multi-sequence-trained models exhibit considerable performance degradation under such suboptimal imaging conditions, BrainVLM maintains diagnostic performance comparable to the full-sequence baseline in scenarios where T2, T2-Flair, or T1 sequences are missing (appendix pp 25-26). A notable exception is the absence of T1c sequences, where we observe a 13.4% decline in F1 score (appendix pp 25-26), highlighting the critical role of post-contrast imaging in brain tumor assessment. Another feature relevant to clinical deployment is that BrainVLM was trained on raw, minimally processed MR images, without skull stripping or segmentation (appendix p 11). Unlike many conventional neuroimaging studies, which typically require preprocessing steps such as skull stripping, BrainVLM directly analyzes raw institutional data, with only slice-level resampling applied (appendix p 11). This minimal preprocessing preserves the complete informational content of the original MRI data, making BrainVLM particularly well-suited for our comprehensive brain tumor characterization study. Collectively, BrainVLM's ability to deliver robust diagnostic performance without preprocessing workflows (e.g., skull stripping) or segmentation requirements, combined with its resilience to missing MRI sequences, establishes its clinical efficacy as an adaptable and practical solution for real-world neuroimaging challenges.

Several limitations exist in our study. First, although online datasets were collected globally, real-world diagnostic applications in this study were confined to East Asian patients. Second, despite improved performance in low-prevalence tumors, data on rare tumors remain relatively limited. Further validation studies with larger sample sizes and more diverse ethnic populations, especially for rare tumor types, are warranted to better evaluate the model's generalizability and robustness. Third, while patient demographics is incorporated into BrainVLM, the model does not currently integrate multimodal clinical data such as medical history, lab findings and neurological examinations. Including these data types could further improve diagnostic accuracy and better align the model with real-world clinical decision-

making. Correspondingly, in the multi-reader study, readers were only given MRI, age, and sex, matching BrainVLM's inputs, to ensure a fair comparison; this may underestimate clinicians' real-world performance, as full clinical context is routinely accessible in routine care. Fourth, although we used patient-level partitioning, independent institutional cohorts, exclusion of recurrent or postoperative follow-up cases, and external test centers that did not contribute to training or validation, residual risks of dataset overlap or data leakage cannot be completely excluded. This is particularly relevant for large public datasets aggregated from multiple sources and because direct cross-center patient matching was not possible after de-identification. Finally, we assessed diagnostic performance, report-generation quality and reader performance with or without AI assistance, but did not examine subsequent treatment decisions and clinical outcomes, such as extent of resection, surgical complications or long-term prognosis.

In conclusion, BrainVLM demonstrates the potential of multimodal foundation models to support complex diagnostic tasks in neuro-oncology. Trained on well-curated multimodal neuroimaging data and validated across internal, multi-center external, and prospective clinical cohorts, the model achieves generalizable performance for classifying a broad spectrum of brain tumors. BrainVLM may serve as a reliable assistant tool to improve the accuracy, efficiency, and consistency of preoperative brain tumor diagnosis and automated report generation. Our framework, which integrates systematic data curation, advanced vision-language modeling, uncertainty quantification, automated report generation, and rigorous clinical evaluation, offers a structured approach for developing specialized foundation models in other medical fields.

Contributors

L.Q. and X.G. conceptualized the study. J.C., Y.W., H.J., L.Q., and A.H. conducted data collection and preprocessing from public repositories. X.G., Z.C., S.K., B.Z., B.W., Z.X., H.Y., Z.W., W.W., B.T., W.H., W.D., W.Z., L.M., A.H., B.L., L.W., Y.C., and W-w.W. collected data from medical centers. X.G., Z.C., S.K., J.C., Y.W., and H.J. performed data preprocessing for institutional data. Y.W., J.C., S.Z., H.J., and L.Q. developed and trained the deep learning algorithms and conducted model analysis. X.G., L.Q., M.I., Y-R.W., F.W., Y.Z., L.Z., and Z.L. provided domain knowledge and interpreted the findings. L.Q., X.G., Y.W., and J.C. drafted the primary manuscript and prepared the figures, with input from all authors. All authors had access to all the data in the study, reviewed and approved the final version of the study. All authors contributed to the general discussion, manuscript revision, and approved the final version. L.Q. and X.G. co-supervised the study. L.Q. and X.G. were responsible for the decision to submit the manuscript.

Data sharing

The online training data are available on their corresponding website. The real-world data are not publicly accessible owing to patient confidentiality requirements. Reasonable requests for access may be considered by the corresponding author, subject to approval from the institutional review boards at all participating centers. The complete code for BrainVLM are available on GitHub at https://github.com/HKU-HealthAI/BrainVLM.

Declaration of interests

We declare no competing interests.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (62306253), the Early Career Fund (27207025, 27204623), the Clinical Research Fund of the National Clinical Research Center for Geriatric Disorders (2023LNJJ19), the Natural Science Foundation of Hunan Province (2023JJ30927), and the Guangdong Natural Science Fund-General Program (2024A1515010233).

Reference

1 Price M, Ballard C, Benedetti J, *et al.* CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2017-2021. *Neuro Oncol* 2024; **26**: vi1–85.

2 Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin* 2023; **73**: 17–48.

3 Miller KD, Ostrom QT, Kruchko C, *et al.* Brain and other central nervous system tumor statistics, 2021. *CA Cancer J Clin* 2021; **71**: 381–406.

4 Louis DN, Perry A, Wesseling P, *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol* 2021; **23**: 1231–51.

5 Karschnia P, Gerritsen JKW, Teske N, *et al.* The oncological role of resection in newly diagnosed diffuse adult-type glioma defined by the WHO 2021 classification: a Review by the RANO resect group. *Lancet Oncol* 2024; **25**: e404–19.

6 Ferreri AJM, Calimeri T, Cwynarski K, *et al.* Primary central nervous system lymphoma. *Nat Rev Dis Primers* 2023; **9**: 29.

7 Frappaz D, Dhall G, Murray MJ, *et al.* EANO, SNO and Euracan consensus review on the current management and future development of intracranial germ cell tumors in adolescents and young adults. *Neuro Oncol* 2022; **24**: 516–27.

8 Horbinski C, Nabors LB, Portnow J, *et al.* NCCN Guidelines® Insights: Central Nervous System Cancers, Version 2.2022. *J Natl Compr Canc Netw* 2023; **21**: 12–20.

9 Zhang K, Zhou R, Adhikarla E, *et al.* A generalist vision-language foundation model for diverse biomedical tasks. *Nat Med* 2024; **30**: 3129–41.

10 Menze BH, Jakab A, Bauer S, *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging* 2015; **34**: 1993–2024.

11 Kim Y, Jeong H, Chen S, *et al.* Medical Hallucinations in Foundation Models and Their Impact on Healthcare. 2025; published online Nov 2. DOI:10.48550/arXiv.2503.05777.

12 Collins GS, Moons KGM, Dhiman P, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024; **385**: e078378.

13 Luo X, Yang Y, Yin S, *et al.* Automated segmentation of brain metastases with deep learning: A multi-center, randomized crossover, multi-reader evaluation study. *Neuro Oncol* 2024; **26**: 2140–51.

14 Wu C, Zhang X, Zhang Y, Hui H, Wang Y, Xie W. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nat Commun* 2025; **16**: 7866.

15 Blankemeier L, Kumar A, Cohen JP, et al. Merlin: a computed tomography vision–language foundation model and dataset. Nature. 2026:1–11.

16 Liu Z, Ning J, Cao Y, *et al.* Video Swin Transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 3192–201.

17 OpenAI. GPT-4; 2023. https://openai.com/research/gpt-4.

18 Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001: 311.

19 Zhao W, Wu C, Zhang X, Zhang Y, Wang Y, Xie W. RaTEScore: A Metric for Radiology Report Generation. 2024; published online Oct 23. DOI:10.48550/arXiv.2406.16845.

20 Delbrouck J-B, Chambon P, Chen Z, *et al.* RadGraph-XL: A Large-Scale Expert-Annotated Dataset for Entity and Relation Extraction from Radiology Reports. In: Findings of the Association for Computational Linguistics ACL 2024. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, 2024: 12902–15.

21 Ström P, Kartasalo K, Olsson H, *et al.* Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol* 2020; **21**: 222–32.

22 Gu B, Desai RJ, Lin KJ, Yang J. Probabilistic medical predictions of large language models. *NPJ Digit Med* 2024; **7**: 367.

23 Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. .

24 Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Mon Wea Rev* 1950; **78**: 1–3.

25 Whitfield BT, Huse JT. Classification of adult-type diffuse gliomas: Impact of the World Health Organization 2021 update. *Brain Pathol* 2022; **32**: e13062.

26 Wu J, Gonzalez Castro LN, Battaglia S, *et al.* Evolving cell states and oncogenic drivers during the progression of IDH-mutant gliomas. *Nat Cancer* 2025; **6**: 145–57.

27 van den Bent MJ, Geurts M, French PJ, *et al.* Primary brain tumours in adults. *Lancet* 2023; **402**: 1564–79.

28 Rao VM, Hla M, Moor M, *et al.* Multimodal generative AI for medical image interpretation. *Nature* 2025; **639**: 888–96.

29 Tanno R, Barrett DGT, Sellergren A, *et al.* Collaboration between clinicians and vision-

558 language models in radiology report generation. *Nat Med* 2025; **31**: 599–608.

559 30 Moor M, Banerjee O, Abad ZSH, *et al.* Foundation models for generalist medical artificial
560 intelligence. *Nature* 2023; **616**: 259–65.

Table 1: Characteristics of our retrospective online dataset, primary dataset, and external test dataset for each brain tumor type.

	Online dataset			**Primary dataset**				**External test dataset**				**Entire dataset**
	No. of subjects	Dimension		No. of subjects	Sex		Age in years Mean ± SD (range)	No. of subjects	Sex		Age in years Mean ± SD (range)	
		2D	3D		Male	Female			Male	Female		
Total	21,993	13,096	8,897	8,813	4,256	4,557	45 ± 18 (1-86)	1,334	665	669	52 ± 15 (1-86)	32,140
MET	1,654	879	775	735	435	300	56 ± 11 (7-79)	116	69	47	60 ± 11 (34-86)	2,505
GCT	81	62	19	378	286	92	19 ± 13 (1-67)	18	11	7	20 ± 13 (1-40)	477
GGN	9,935	4,743	5,192	3,028	1,681	1,347	41 ± 19 (1-86)	414	244	170	50 ± 15 (5-79)	13,377
MEN	5,363	3,804	1,559	2,395	723	1,672	53 ± 12 (1-81)	355	115	240	56 ± 13 (8-79)	8,113
TSR	674	614	60	796	402	394	46 ± 16 (2-81)	186	98	88	51 ± 15 (5-75)	1,656
MNM	912	457	455	447	234	213	39 ± 18 (1-78)	76	40	36	43 ± 16 (8-76)	1,435
CPN	1,865	1,392	473	416	138	278	46 ± 15 (1-75)	90	43	47	54 ± 10 (34-75)	2,371
CPT	259	216	43	129	63	66	28 ± 20 (1-66)	6	1	5	45 ± 10 (39-62)	394
HEM	535	396	139	164	100	64	47 ± 22 (2-77)	44	27	17	62 ± 11 (30-82)	743
EMB	288	165	123	205	131	74	18 ± 15 (1-70)	18	12	6	18 ± 15 (3-40)	511
PIN	346	312	34	70	36	34	36 ± 20 (1-73)	6	2	4	39 ± 13 (21-57)	422
MEL	81	56	25	50	27	23	50 ± 14 (14-73)	5	3	2	36 ± 25 (21-69)	137

Abbreviations: MET = brain metastases, GCT = germ cell tumors, GGN = gliomas, glioneuronal tumors, and neuronal tumors, MEN = meningioma, TSR = tumors of the sellar region, MNM = mesenchymal, non-meningothelial tumors, CPN = cranial and paraspinal nerve tumors, CPT = choroid plexus tumors, HEM = hematolymphoid tumors, EMB = embryonal tumors, PIN = pineal region tumors and MEL = melanocytic tumors.

Table 2: Comparative diagnostic performance across neuroradiologists-only, BrainVLM, and AI-augmented neuroradiologists paradigms. Twelve neuroradiologists with varying years of experience (junior [3-5 years], senior [5-10 years], expert [10+ years]) participated in this blinded crossover study. Boldface indicates the highest performance metric within each tumor type across AI, neuroradiologists, and AI-augmented neuroradiologists cohorts.

	F_1 score						
	AI model	Junior	Junior-AI	Senior	Senior-AI	Expert	Expert-AI
MET	0.74	0.51	0.62	0.44	0.66	0.80	**0.87**
GCT	0.75	0.28	0.54	0.40	0.59	0.70	**0.79**
GGN	0.83	0.64	0.72	0.66	0.80	0.84	**0.91**
MEN	0.85	0.74	0.81	0.77	0.85	**0.90**	**0.90**
TSR	0.86	0.74	0.77	0.78	0.81	0.85	**0.88**
MNM	0.56	0.10	0.33	0.07	0.42	0.50	**0.58**
CPN	0.77	0.57	0.74	0.69	0.79	0.73	**0.84**
CPT	0.65	0.30	0.53	0.36	0.68	0.77	**0.88**
HEM	0.75	0.22	0.56	0.21	0.70	0.70	**0.86**
EMB	**0.82**	0.14	0.62	0.31	0.75	0.54	0.68
PIN	0.77	0.43	0.62	0.48	0.68	0.70	**0.75**
MEL	0.25	0.00	0.40	0.00	0.69	0.63	**0.80**
Frequency -weighted F_1	0.78 [0.73, 0.83]	0.52 [0.48, 0.55]	0.67 [0.63, 0.71]	0.55 [0.51, 0.58]	0.74 [0.70, 0.77]	0.78 [0.74, 0.82]	**0.85 [0.81, 0.89]**
Accuracy	0.79 [0.74, 0.83]	0.55 [0.52, 0.59]	0.69 [0.65, 0.72]	0.59 [0.56, 0.62]	0.75 [0.72, 0.78]	0.80 [0.75, 0.83]	**0.86 [0.82, 0.89]**
Total Time (Mean ± SD)	**51** ± **3**	189 ± 41	112 ± 13	161 ± 43	104 ± 21	100 ± 5	75 ± 15

Abbreviations: MET = brain metastases, GCT = germ cell tumors, GGN = gliomas, glioneuronal tumors, and neuronal tumors, MEN = meningioma, TSR = tumors of the sellar region, MNM = mesenchymal, non-meningothelial tumors, CPN = cranial and paraspinal nerve tumors, CPT = choroid plexus tumors, HEM = hematolymphoid tumors, EMB = embryonal tumors, PIN = pineal region tumors and MEL = melanocytic tumors. SD: Standard Deviation.

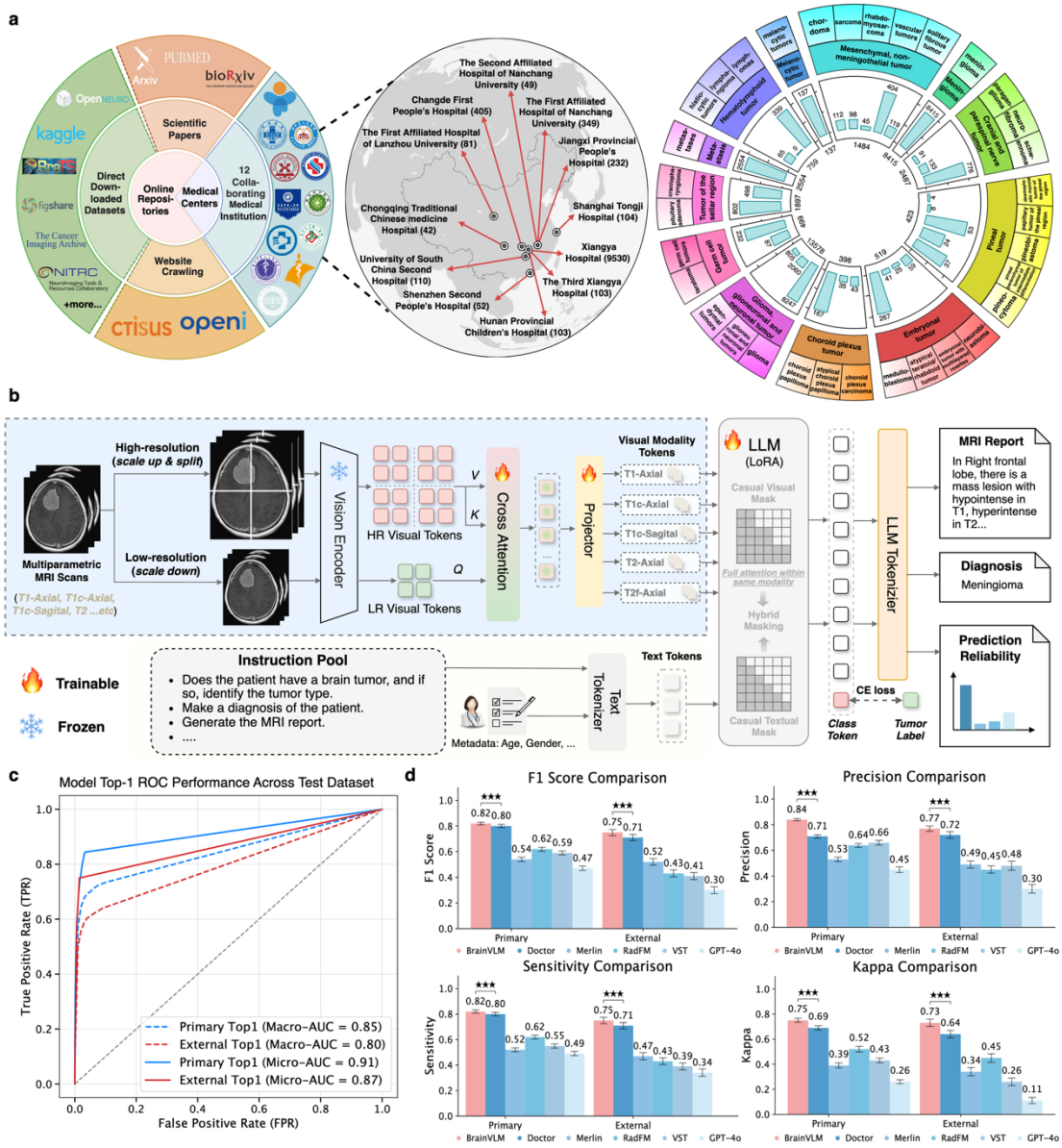

Figure 1: **Overview of this study. a,** A multi-modal brain tumor dataset was constructed from 39 online sources and 12 collaborating medical institutions, comprising 47,947 patients. Each case includes either 2D MRI slices or multi-sequence 3D MRI scans (T1-weighted[T1], T1 contrast-enhanced [T1c], T2-weighted [T2], T2-Flair [T2f]), paired with pathologically confirmed diagnoses, optional radiology reports (or text descriptions). The detailed number of patients across 12 brain tumor types and their subtypes is shown on the right panel. **b,** we developed BrainVLM, a vision-language foundation model designed for brain tumor analysis. The model processes multi-parametric MRI scans at different resolutions through a shared vision encoder and integrates text data via an LLM-based architecture. BrainVLM was trained to classify tumor types with quantified diagnostic uncertainty and a radiology report. **"Frozen"** components (Vision Encoder and LLM backbone) retain their pre-trained weights to ensure structural stability and leverage broad medical knowledge. **"Trainable"** components (MLP projector and LoRA-based adapters) are optimized during fine-tuning via backpropagation. **c,** ROC curves for top-1 predictions on the retrospective primary and external datasets, with macro-averaged AUCs of 0.85/0.80 and micro-averaged 0.91/0.87, respectively. **d,** BrainVLM outperformed neuroradiologists and baseline models in sensitivity, precision, F1 score, and Cohen's κ across both datasets. Two-tailed Wilcoxon signed-rank tests were performed to compare BrainVLM with neuroradiologists for each metric. P values were not adjusted for multiple comparisons because each metric was prespecified and interpreted as a distinct measure of diagnostic performance. Statistical significance is indicated as ns (not significant, $p>0.05$), ⋆ ($p<0.05$), ⋆⋆ ($p<0.01$), and ⋆⋆⋆ ($p<0.0001$). Bootstrap resampling (1,000 replicates) was used to estimate 95% confidence intervals, shown as error bars around the mean.

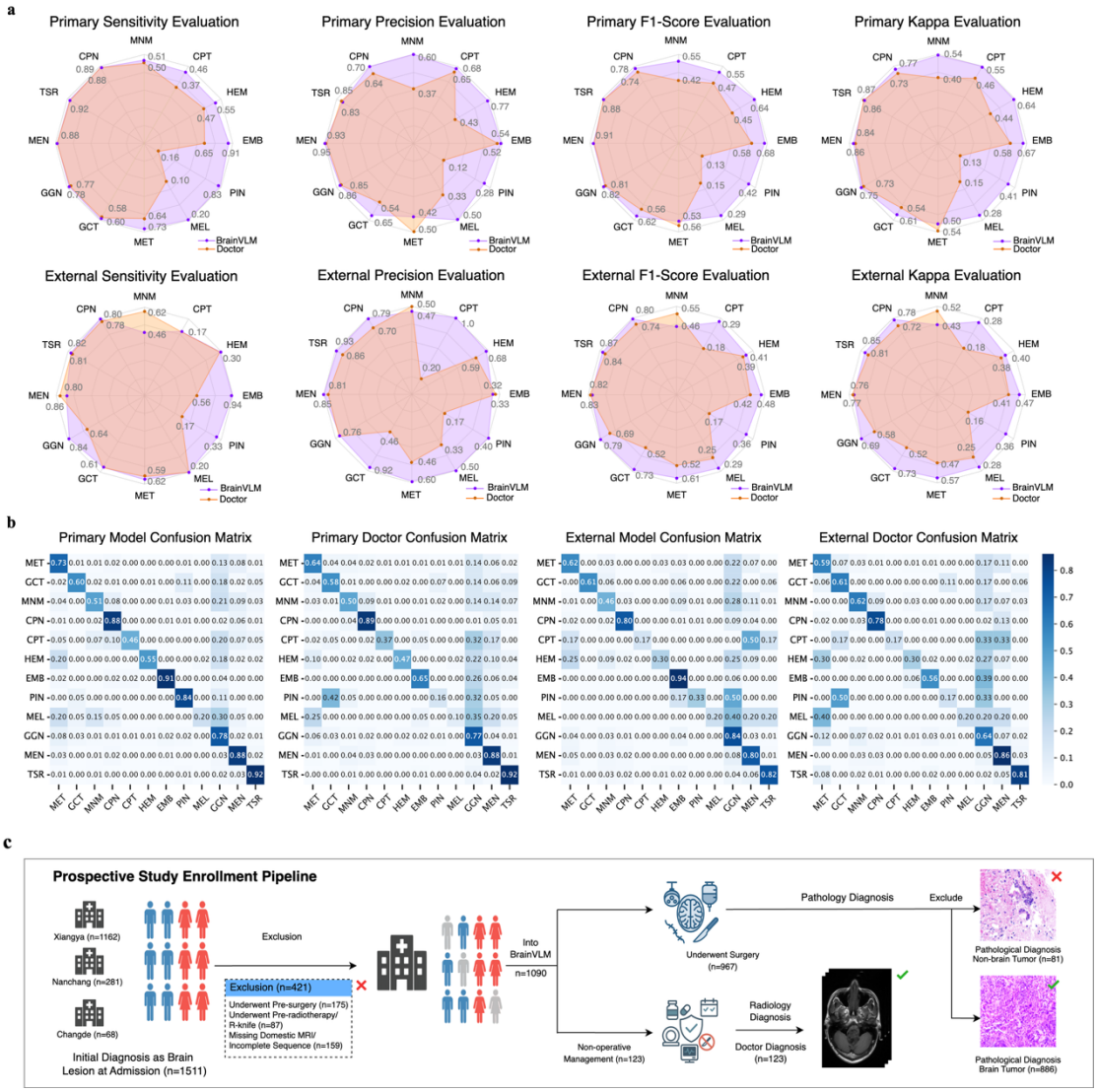

Figure 2 a, Performance comparison of BrainVLM and board-certified neuroradiologists on 12 tumor types across primary and external datasets. Model and human expert results are visualized as a dodecagon, with each vertex corresponding to a specific tumor type. For both datasets, we report sensitivity, precision, F1 score, and Cohen's κ for each class. BrainVLM demonstrates either superior or comparable performance relative to neuroradiologists on both primary and external cohorts, including for rare tumor types (e.g., EMB, GCT, HEM). **b,** Confusion matrices comparing BrainVLM's predictions to those of neuroradiologists across 12 tumor types in both primary and external datasets. The foundation model showed stronger diagonal concentration than neuroradiologists, indicating improved classification accuracy. Each matrix was normalized by true labels to highlight per-class performance. Notably, the model's diagnostic error patterns showed substantial concordance with those of human experts. **c,** Flowchart of prospective patient enrollment. A total of 1,511 patients with suspected brain lesions were prospectively enrolled. After excluding those who had received prior treatment or lacked diagnostic sequences, 1,090 patients were eligible for evaluation. Among them, 967 underwent surgery; after excluding 123 cases with non-brain tumor pathology, 886 patients with confirmed brain tumors were included for model validation. The remaining 81 patients who received non-operative management were also included for evaluation.

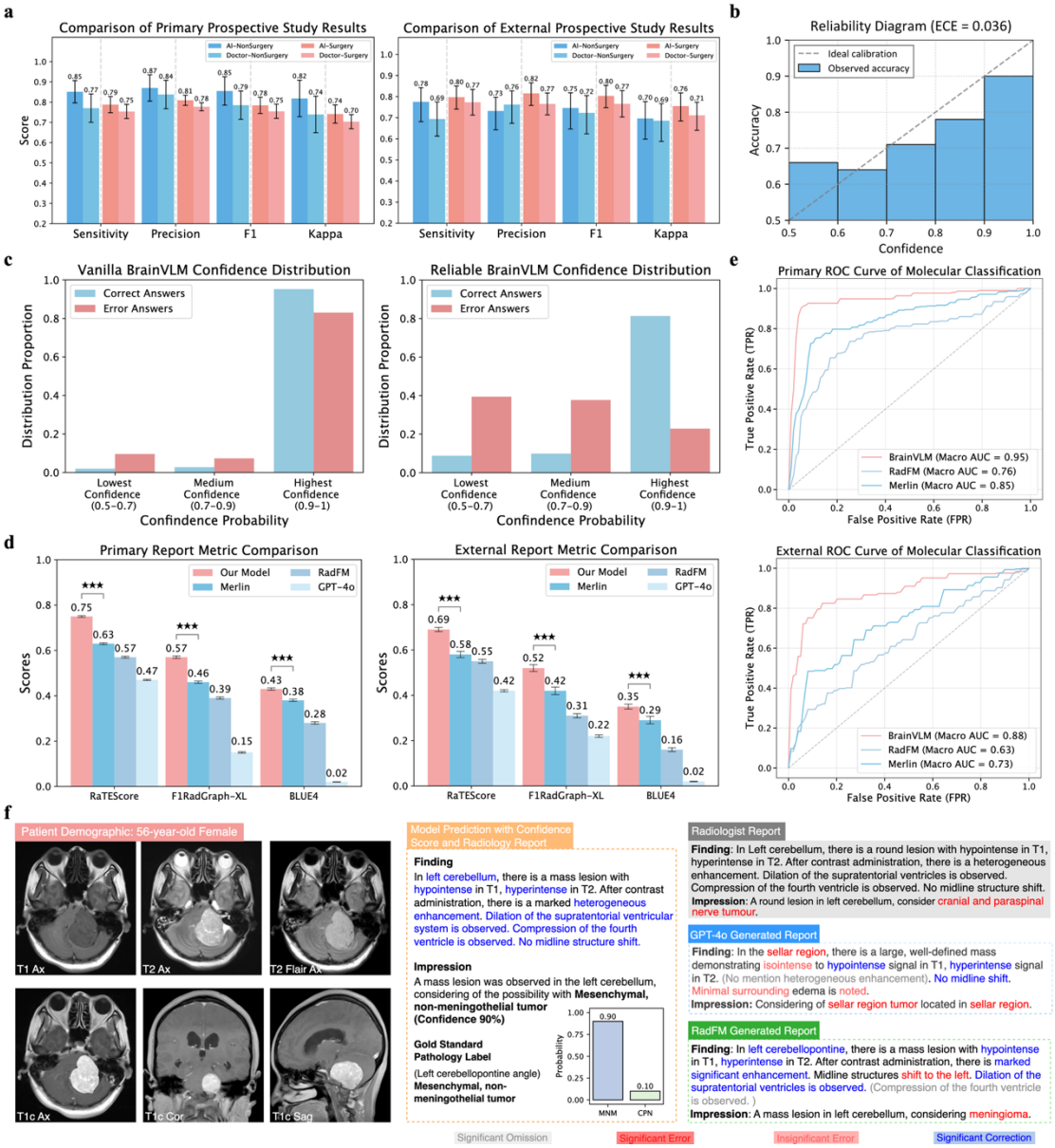

Figure 3: Performance comparison in the prospective cohort, as well as uncertainty quantification, comparative evaluation of radiology report generation, and molecular subgroup prediction. a, Comparison of primary and external prospective study results. The bar charts evaluate BrainVLM (AI) versus human doctors across surgical and non-surgery groups using Sensitivity, Precision, F1, and Kappa metrics. **b,** Reliability diagram (ECE = 0.036) illustrating model calibration and the alignment between observed accuracy and confidence probability. **c,** the comparison of confidence score distributions between Vanilla BrainVLM, which does not incorporate our consensus-driven strategy for uncertainty quantification, and Reliable BrainVLM, which benefits from the integration of this strategy. **d,** Report metric comparison of four models using RaTEScore, F1RadGraph-XL, and BLEU4 on both primary and external datasets. Two-tailed Wilcoxon signed-rank tests were performed to compare BrainVLM with comparator models; p values were not adjusted for multiple comparisons, as each metric was interpreted individually rather than as repeated tests of a single hypothesis. **e,** ROC curves comparison for preoperative molecular subgroup prediction in adult-type diffuse gliomas among RadFM, Merlin, and our model on both primary and external datasets. **f**, A representative case study comparing generated reports from our model, ChatGPT-4o, and RadFM against the human radiologist's reference report, highlighting significant corrections, errors, and omissions. For BrainVLM's diagnosis, the associated probability and confidence score are also presented.

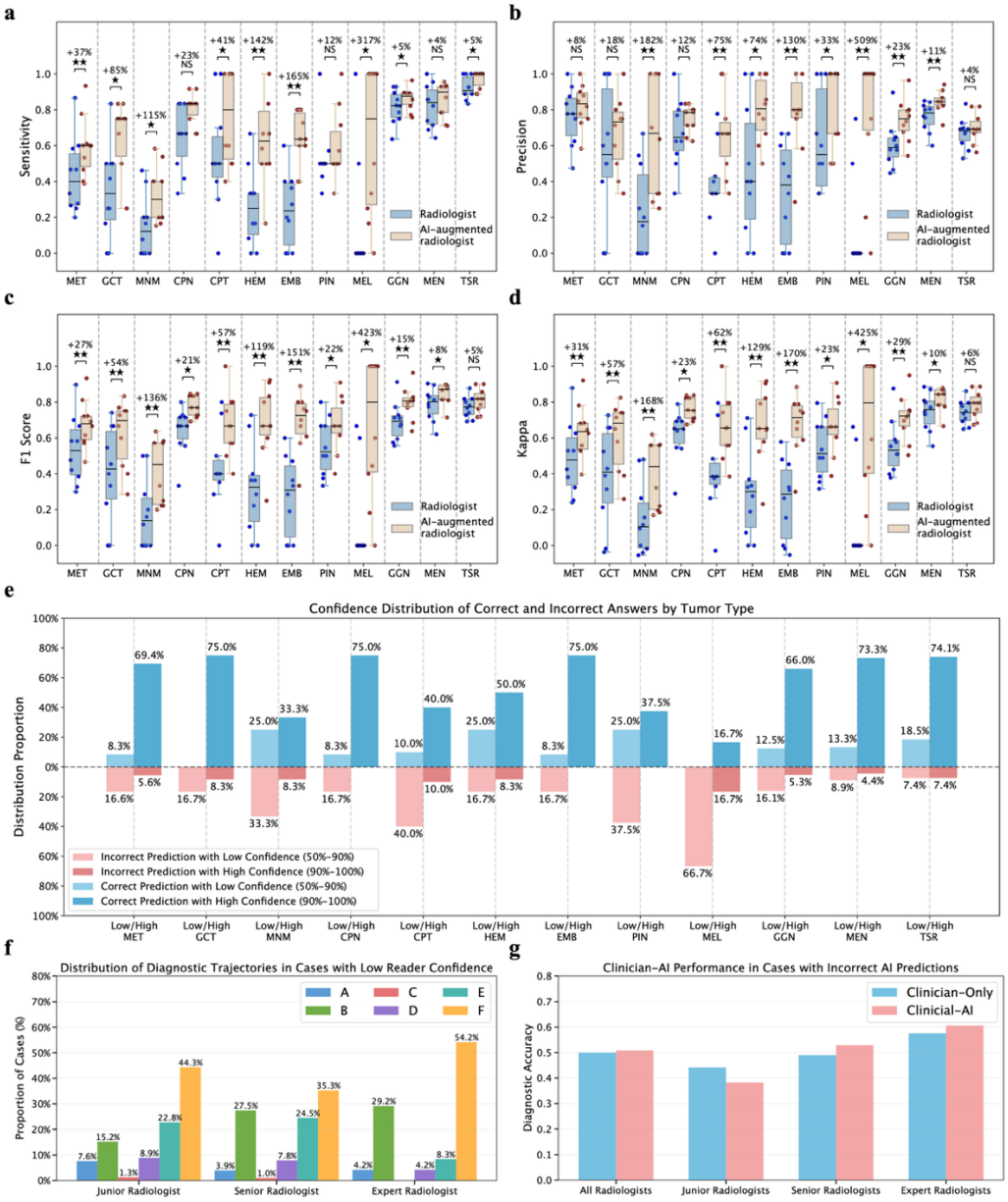

Figure 4: **Blinded multi-reader study on 248 patients with interpretations by 12 neuroradiologists (5 juniors [3-5 years], 4 seniors [5-10 years], 3 experts [10+ years]) with and without AI assistance. a-d,** Comparison between the diagnostic performance of neuroradiologists and AI-augmented neuroradiologists. Box plots summarize sensitivity, precision, F1 score, and Cohen's κ for individual neuroradiologists and their AI-augmented counterparts. Pairwise comparisons between each neuroradiologist and their corresponding AI-augmented counterpart were performed using two-tailed Wilcoxon signed-rank tests without correction for multiple comparisons. Statistical significance is indicated as ns (not significant, $p>0.05$), ⋆ ($p<0.05$), ⋆⋆ ($p<0.01$), and ⋆⋆⋆ ($p<0.001$). P values were not adjusted for multiple comparisons because per-tumor-type analyses were prespecified and interpreted separately for each tumor category. **e,** Confidence score distribution for BrainVLM's predictions on these 248 cases across 12 major WHO CNS5 brain tumor types. The figure is vertically divided: upper blue bars show confidence distributions for correct predictions, lower pink bars show incorrect ones. **f,** Distribution of Diagnostic Trajectories in Cases with Low Reader Confidence. The bar chart shows the distribution of four diagnostic trajectories after exposure to BrainVLM among the subset with the lowest 25% of initial clinician confidence, stratified by reader seniority. Trajectories are defined as: (A) maintaining a correct diagnosis despite incorrect AI; (B) successfully correcting an initial error; (C) erroneously changing a correct diagnosis due to incorrect AI; (D) both clinician and AI remaining incorrect; (E) failing to adopt a correct AI suggestion; and (F) both initial clinician diagnosis and AI suggestion being correct, with the final diagnosis remaining correct. **g,** Clinician-AI performance in cases with incorrect AI predictions. The bar chart compares clinician diagnostic accuracy without AI assistance and after review of BrainVLM output, restricted to cases in which BrainVLM generated an incorrect diagnosis, stratified by reader seniority (junior, senior, and expert).

Click here to access/download

Supplementary Materials

BrainVLM supplementary appendix