DEEPSITE-MRI: A PUBLIC BENCHMARK FOR DEEP-SEATED INTRACRANIAL TUMOR DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep-seated intracranial tumors (thalamus, basal ganglia, brainstem) are adjacent to critical functional areas, making direct surgical resection typically infeasible; clinicians rely on stereotactic biopsy to determine subtypes for treatment planning. However, existing brain tumor datasets focus on superficial tumors, and no public benchmark with biopsy-confirmed labels exists for deep-seated tumors. To bridge this gap, we present **DeepSite-MRI**, the first such benchmark, comprising 249 biopsy-confirmed cases across four subtypes (Glioma, PCNSL, Germinoma, Rare), together with a **Two-Stage spatial Prior pipeline (TSP)** that combines VLM-guided coarse localization with MLP-based refinement, reducing per-case annotation time from approximately 90 minutes to approximately 3 minutes (96% clinical acceptance rate); this paradigm is transferable to other data-scarce medical imaging domains. Meanwhile, we systematically evaluate 10 mainstream architectures spanning CNNs and Transformers to characterize the difficulty and research value of this benchmark. Results show that even well-established methods that perform strongly on large-scale brain tumor benchmarks underperform on DeepSite-MRI (best accuracy 78.90%), which not only demonstrates the potential for further research in this direction, but also indicates that the benchmark can serve as a foundation for data-constrained learning and clinical translation. Code and data will be released upon acceptance.

1 INTRODUCTION

Brain tumors are among the most common malignant diseases of the central nervous system, posing a severe threat to patients' survival and quality of life Price et al. (2025). Timely and accurate treatment can significantly extend survival and improve prognosis. Treatment strategies for brain tumors can be broadly categorized into two types based on anatomical location. (i) For superficial tumors (located in the cerebral cortex or subcortical regions), surgical resection is typically the primary clinical intervention. (ii) For deep-seated tumors (located in core functional areas such as the thalamus, basal ganglia, and brainstem), direct resection is extremely risky and often unfeasible due to their proximity to critical neural nuclei controlling motor, language, and vital functions. Instead, clinicians typically perform a stereotactic biopsy first to determine the histopathological subtype, and then administer targeted treatment accordingly. For instance, Primary Central Nervous System Lymphoma (PCNSL) primarily requires chemotherapy Saad et al. (2023), whereas Glioblastoma (GBM) necessitates surgery combined with radiochemotherapy Zhang et al. (2019). However, the biopsy procedure itself carries risks of complications such as hemorrhage and infection, and the sampling results are inherently limited by target selection, leading to inconclusive or even misleading pathological results Fawaz et al. (2025). This clinical bottleneck motivates us to pursue reliable prediction of pathological subtypes from preoperative MRI alone Holtkamp et al. (2025), thereby providing a vital reference for both clinical decision-making and biopsy target planning.

Recent advances in deep learning and Vision-Language Models (VLMs) have made it technically feasible to predict pathological subtypes from preoperative MRI alone. However, existing brain tumor datasets (*e.g.*, BraTS Menze et al. (2015); Baid et al. (2021), TCGA Bakas et al. (2017)) do not support this task, as they primarily consist of superficial or hemispheric tumors whose imaging characteristics differ substantially from deep-seated lesions (*e.g.*, deep-seated lesions are typically smaller and exhibit blurrier boundaries), making them inadequate substitutes. To date, no public benchmark with biopsy-confirmed labels exists for deep-seated intracranial tumors, making the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

construction of such a specialized dataset an essential yet particularly challenging task in this field. Specifically, constructing such a benchmark involves two major practical challenges:

`Challenge-Collection`: Constructing this benchmark requires preoperative MRI scans paired with biopsy-confirmed pathological labels. However, deep-seated intracranial tumors are relatively rare, and among these cases, only a subset of patients undergo stereotactic biopsy to obtain a postoperative pathological report, further narrowing the pool of eligible samples. `Challenge-Annotation`: Furthermore, to construct a dataset that supports downstream research such as tumor localization and classification, fine-grained spatial annotations delineating the tumor region are required. This requires specialized expertise, and manually labeling small, indistinct lesions on 3D MRI can take an experienced physician up to 90 minutes per case.

Overcoming these challenges, this paper presents **DeepSite-MRI**, the first public 3D MRI benchmark for deep-seated intracranial tumors with biopsy-confirmed pathological labels, comprising 249 cases across four subtypes (Glioma, PCNSL, Germinoma, Rare) collected from multiple neurosurgical centers. To alleviate the annotation burden on physicians, we design a Two-Stage spatial prior labeling Pipeline (TSP), a fully automatic annotation tool that combines VLM-based localization with fine-grained refinement, reducing per-case annotation time from up to 90 minutes to approximately 3 minutes. All annotations are rigorously reviewed and confirmed by board-certified physicians to ensure clinical-grade reliability. Built upon DeepSite-MRI, we benchmark 10 mainstream architectures widely used in brain tumor analysis. The best-performing model achieves only 78.90% accuracy on the four-class classification task, highlighting both the methodological gap in existing approaches and the necessity of constructing such a dedicated benchmark to drive future advances.

Our contributions are summarized as follows: **(i) Dataset:** We construct and release DeepSite-MRI, the first public benchmark for deep-seated intracranial tumors with biopsy-confirmed pathology, filling the critical data gap for this challenging clinical task. **(ii) Annotation Tool:** We propose and open-source the TSP two-stage labeling pipeline, which compresses the cost of generating spatial priors for a single case from hours to minutes. Following expert evaluation, this pipeline provides a potentially reusable approach for constructing similar datasets in data-scarce medical domains. **(iii) Benchmark Insights:** Through systematic experiments, we reveal the core challenges of this task and provide valuable quantitative references for future method design and development, supporting downstream research directions including annotation methodology transfer, long-tailed recognition, data-efficient transfer learning, and clinical translational applications.

2 Related Work

Representative datasets include BraTS Menze et al. (2015), BraTS 2021 Baid et al. (2021), BraTS-MEN-RT LaBella et al. (2026), BraTS-METS Moawad et al. (2023), BraTS-PEDs Kazerooni et al. (2023a), TCGA-GBM/LGG Bakas et al. (2017), MSD Antonelli et al. (2022), CE-MRI Figshare Cheng (2017), Kaggle BTC Nickparvar (2021), Mendeley Nickparvar (2024), UCSF-BMSR Rudie et al. (2024), CBTN Kazerooni et al. (2023b), as well as private single-center collections on PCNSL Fu et al. (2025) and germinoma Supbumrung et al. (2023). We review these datasets along three dimensions (detailed comparison in Tab. 1): lesion anatomical coverage, label source, and annotation methodology.

Lesion anatomical coverage. Existing public datasets primarily cover hemispheric gliomas, meningeal/skull-base tumors, and brain metastases. In contrast, primary tumors of the thalamus, basal ganglia, and brainstem remain largely absent from public benchmarks, while the few relevant PCNSL and germinoma datasets are private and cannot support independent replication or systematic cross-method evaluation. DeepSite-MRI addresses this gap with 249 deep-seated tumor cases spanning four pathological subtypes: Glioma, PCNSL, Germinoma, and Rare.

Source of tumor labels. Most existing datasets assign tumor subtypes from radiological assessment of MRI appearance, whereas the available PCNSL and germinoma collections use stereotactic biopsy-confirmed pathology. No existing public dataset provides biopsy-confirmed labels for deep-seated brain tumors. DeepSite-MRI provides stereotactic biopsy-confirmed pathological labels for all 249 cases, establishing a public pathological benchmark for this anatomical region.

Annotation methodology. Existing datasets with spatial annotations primarily rely on expert manual delineation, while several 2D classification datasets provide only image-level labels. For deep-seated tumors, indistinct lesion boundaries and limited reference cases make manual annotation

particularly costly, requiring up to 90 minutes per case. DeepSite-MRI instead adopts TSP, a two-stage pipeline combining VLM-guided coarse localization with subject-adaptive refinement, reducing per-case annotation time to approximately 3 minutes with expert review.

Table 1: Comparison of brain tumor MRI datasets across six dimensions. **Location abbreviations:** H.Glio. = hemispheric glioma; Men. = meningioma; Pit. = pituitary adenoma; B.Meta. = brain metastasis; Ped. = pediatric; P.F. = posterior fossa; Mid. = midline; DMG = diffuse midline glioma; Th. = thalamus; B.G. = basal ganglia; B.S. = brainstem. **Task:** Cls. = classification; Seg. = segmentation; Det. = detection. **Annot.:** Expert manual = voxel-level delineation by expert physicians; Expert reading = image-level label by radiologist reading (no spatial annotation); TSP + Expert review = automatic annotation via the Two-Stage spatial Prior (TSP) pipeline, verified by board-certified physicians. ✓ satisfied; △ partially/restricted; ✗ not satisfied. †CBTN requires an approved data-access agreement (△); BraTS-PEDs and CBTN partially involve midline structures but are confined to pediatric subtypes not transferable to the adult deep-seated setting (△).

Dataset	Tumor Location (abbreviated)	Dim.	Cases	Task	Annot.	Biopsy-confirmed	Public
BraTS Menze et al. (2015); Baid et al. (2021)	H.Glio.	3D	1,251	Seg.	Expert manual	✗	✓
TCGA-GBM/LGG Bakas et al. (2017)	H.Glio.	3D	461	Seg.	Expert manual	✗	✓
MSD Task01 Antonelli et al. (2022)	H.Glio.	3D	750	Seg.	Expert manual	✗	✓
BraTS-MEN-RT LaBella et al. (2026)	Men.	3D	570	Seg.	Expert manual	✗	✓
CE-MRI Figshare Cheng (2017)	Men./Pit./Glio.	2D	3,064	Cls.	Expert reading	✗	✓
Kaggle BTC Nickparvar (2021)	Men./Pit./Glio.	2D	7,023	Cls.	Expert reading	✗	✓
Mendeley Nickparvar (2024)	Men./Pit./Glio.	2D	11,148	Cls.	Expert reading	✗	✓
UCSF-BMSR Rudie et al. (2024)	B.Meta.	3D	412	Det.+Seg.	Expert manual	✗	✓
BraTS-METS Moawad et al. (2023)	B.Meta.	3D	238	Seg.	Expert manual	✗	✓
CBTN Kazerooni et al. (2023b)	Ped.(P.F./Mid.)	3D	4,900+	Cls.+Seg.	Expert manual	✗	△
BraTS-PEDs Kazerooni et al. (2023a)	Ped. DMG	3D	228	Seg.	Expert manual	✗	✓
PCNSL study Fu et al. (2025)	Deep.(Th./B.G.)	3D	135	Seg.	Expert manual	✓	✗
Germinoma study Supbumrung et al. (2023)	Deep.(Th./B.G.)	3D	38	Cls.	Expert manual	✓	✗
DeepSite-MRI (Ours)	**Deep.(Th./B.G./B.S.)**	**3D**	**249**	**Cls.+Seg.**	**TSP + Expert review**	✓	✓

3 DeepSite-MRI Dataset

3.1 Data Acquisition and Open Standards

Clinical Source and Ethical Standards: To overcome `Challenge-Collection`, DeepSite-MRI was prospectively collected in a multicenter study from major tertiary neurosurgical centers, one of which is among the largest neurosurgical centers in China, performing over 500 stereotactic biopsies annually, providing a sufficient pool of biopsy-confirmed deep-seated tumor cases. The study protocol was approved by the Institutional Review Board (IRB) of the authors' affiliations. To protect patient privacy, all raw data underwent de-identification in a secure local environment, stripping all protected health information (PHI) and facial features. Details are shown in Fig. 1.

Inclusion and Exclusion Criteria: To ensure the dataset's specificity for deep brain tumor assessment and the purity of image quality, this cohort adhered to the following strict criteria. *Inclusion criteria:* (i) Imaging suggested a deep intracranial space-occupying lesion, assessed by a multidisciplinary team (MDT) as having an excessively high risk for open surgical resection, leading to a clinical decision for stereotactic biopsy; (ii) The biopsy was successfully completed, and the postoperative histopathological report definitively diagnosed Glioma, Primary Central Nervous System Lymphoma (PCNSL), Germinoma, or other rare pathologies (Rare); (iii) A structurally complete preoperative T1-weighted contrast-enhanced (T1-CE) sequence without severe motion artifacts was available. *Exclusion criteria:* Pathologically confirmed benign or non-neoplastic lesions (e.g., cavernous hemangioma, inflammatory pseudotumor); missing core sequences or image registration failures; and cases where patients/families refused authorization for data usage.

Cohort Demographics and Pathological Distribution: The final cohort comprises 249 patients, including 135 males (54.2%) and 114 females (45.8%), aged 4–86 years (mean 43.7 ± 23.2). The four pathological classes exhibit a natural imbalance: Glioma (149, 59.8%), PCNSL (72, 28.9%),

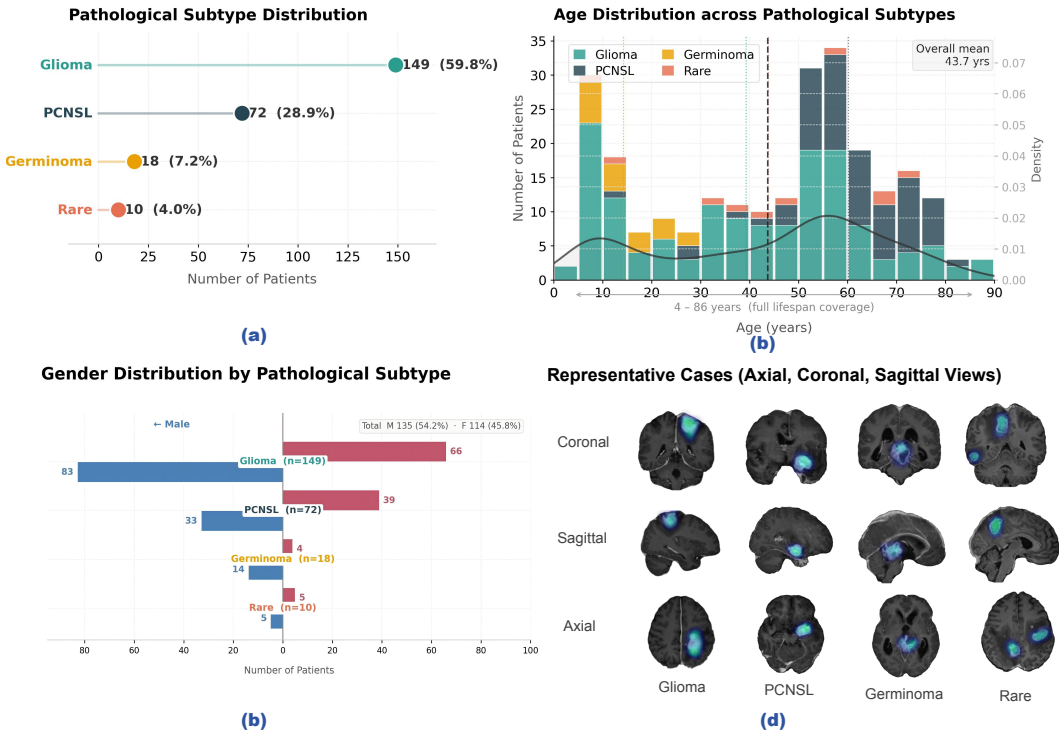


Figure 1: Overview of the DeepSite-MRI dataset. **(a)** Pathological subtype distribution across 249 cases, reflecting real-world incidence rates with natural class imbalance. **(b)** Age distribution of the cohort, demonstrating full lifespan coverage from pediatric to elderly patients. **(c)** Gender distribution highlighting the demographic breakdown across different pathological subtypes. **(d)** Representative MRI cases for each subtype displayed in axial, coronal, and sagittal views. Tumor regions (highlighted) are located in deep functional areas including the thalamus, basal ganglia, and brainstem, illustrating the characteristic small lesion size, low signal contrast, and complex anatomical background that define the core difficulty of this benchmark.

Germinoma (18, 7.2%), and Rare (10, 4.0%). This distribution is retained in the benchmark to reflect the clinical setting.

Scanner Diversity. DeepSite-MRI was collected across multiple centers and was acquired using multiple MRI systems from three manufacturers (Philips, Siemens, and GE). Among the 249 cases with complete scanner information, four MRI systems were represented. Most cases were acquired on the Philips Ingenia CX (232/249), with the remaining cases acquired on Siemens Prisma, GE SIGNA Premier, and Philips Ingenia systems. Thus, the benchmark incorporates acquisition variability across scanners, manufacturers, and clinical sites, although the strong concentration on one scanner limits the extent to which this heterogeneity can be interpreted as comprehensive external-center validation. The complete scanner-by-pathology distribution is provided in Appendix G.

Rationale for Modality Selection. We selected T1-CE as the sole input modality because it is a core sequence in the preoperative assessment of deep intracranial tumors and is more widely available in routine clinical workflows than advanced sequences such as DWI or MRS. This choice improves clinical deployability and cross-institutional reproducibility while isolating algorithmic differences from gains due to additional modalities.

3.2 A Two-Stage Scalable Spatial Prior Annotation Pipeline

Core Challenge and Motivation. To overcome `Challenge-Annotate`, we design TSP, a fully automatic two-stage annotation pipeline that compresses per-case annotation time from up to 90 minutes to approximately 3 minutes, requiring no pre-trained classification model (unlike

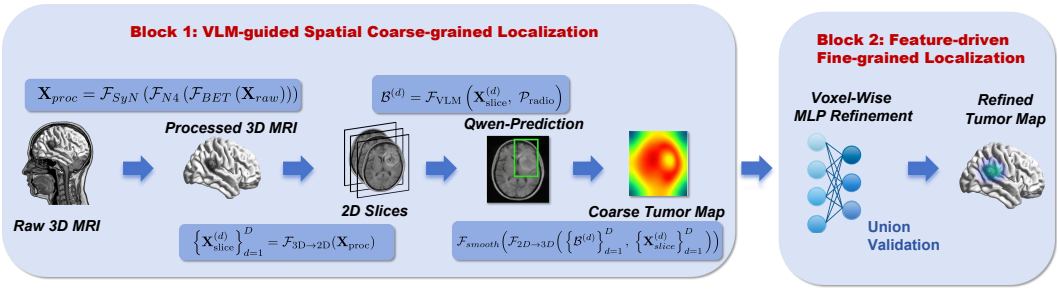


Figure 2: Overview of the Two-Stage Scalable Spatial Prior annotation Pipeline (TSP). **Block 1** performs VLM-guided spatial coarse-grained localization: standardized T1-CE volumes are sliced axially and fed into Qwen3-VL-Thinking with a structured radiological prompt to generate per-slice bounding boxes, which are reconstructed and smoothed into a 3D coarse prior map $\mathbf{X}_{\text{tumor-coarse}}$. **Block 2** performs feature-driven fine-grained refinement: a subject-adaptive MLP trained on high-confidence voxels produces a refined prior $\mathbf{X}_{\text{tumor-refined}}$, which is concatenated with the original T1-CE volume as an additional input channel to the downstream classification backbone.

GradCAM/CAM-based methods Selvaraju et al. (2020)) and no manual point or box prompts (unlike SAM/SAM2-based approaches Cheng et al. (2023)). Expert effort is reserved solely for final quality validation. TSP operates under a single-lesion assumption: each scan is presumed to contain one spatially contiguous dominant lesion.

Block 1: VLM-guided Coarse Localization. Each T1-CE volume is preprocessed through the standard FreeSurfer (Fischl, 2012), then sliced axially into 2D frames. Each slice $\mathbf{X}_{\text{slice}}^{(d)}$ is fed to Qwen3-VL-Thinking (selected from several candidate VLMs; see **Appendix C** for comparisons) with a structured radiological prompt $\mathcal{P}_{\text{radio}}$ that directs the model to identify abnormal enhancement patterns, edema gradients, and midline asymmetry, yielding per-slice bounding box predictions:

$$\mathcal{B}^{(d)} = \mathcal{F}_{\text{VLM}}\left(\mathbf{X}_{\text{slice}}^{(d)}, \mathcal{P}_{\text{radio}}\right), \tag{1}$$

where $\mathcal{B}^{(d)} = \emptyset$ if no credible lesion evidence is identified. Valid bounding boxes are stacked and sequentially smoothed via Gaussian filtering ($\sigma = 1.5$ voxels) and morphological dilation (radius $= 3$ voxels) to form a continuous 3D coarse prior $\mathbf{X}_{\text{tumor-coarse}}$, maximally covering the full tumor extent while suppressing stair-step artifacts from per-slice discretization.

Block 2: Feature-driven Fine-grained Refinement. The coarse prior tends to overestimate lesion extent by including surrounding edema. We therefore train a lightweight voxel-wise MLP independently for each subject using high-confidence coarse-prior voxels, with local MRI intensity, gradient, texture, and prior confidence as input features. The refined prior is obtained by the union of the MLP prediction and the coarse prior:

$$\mathbf{X} * \text{tumor-refined} = \mathbf{X} * \text{tumor-coarse} \cup \mathcal{F} * \text{MLP!}\left(\mathbf{X} * \text{img} \mid \mathbf{X}_{\text{tumor-coarse}} > \tau\right). \tag{2}$$

The union preserves candidate tumor regions that may be missed by the MLP. The resulting prior is concatenated with the original T1-CE volume as an additional input channel; implementation details are provided in **Appendix A**.

3.3 Expert Validation Study

Validation Protocol. To quantitatively assess the clinical reliability of TSP annotations, we conducted an independent double-blind expert study. The study was approved by the institutional ethics committee and carried out independently from the data acquisition process. The evaluation was performed by three board-certified neurosurgeons from independent tertiary hospitals, with no affiliation to the data collection center. All experts held at least associate chief physician titles, with over 10 years of experience in stereotactic biopsy for intracranial tumors, and an annual case volume exceeding 50. All assessments were conducted independently; experts were blinded to both the annotation source (manual vs. automatic) and the ratings of other reviewers. We constructed a validation subset of 50 cases (approximately 20%) from the full cohort of 249 patients using stratified random sampling, ensuring that the distribution of tumor types was preserved.

Evaluation Setup. For each clinical case, a comprehensive standardized evaluation package was provided, including: (i) raw MRI scans (axial, coronal, and sagittal views) without any annotation overlays; (ii) TSP-generated spatial priors, displayed separately as semi-transparent colored contours. The two components were presented sequentially on independent display screens, with randomized order to minimize potential ordering effects. Each case was evaluated on two primary dimensions: a five-point Accuracy Score (AS, where AS ≥ 4 denotes acceptable localization) and a binary Clinical Acceptability rating (CA: yes/no, indicating whether the annotation is usable without further correction). Detailed display rubrics are provided in Tab. 3 of **Appendix C**

Expert-TSP Agreement Results. Among the 50 evaluated cases, 96% (48/50) of TSP annotations were deemed clinically acceptable by the expert panel. The overall mean Accuracy Score was 4.28 ± 0.75 (out of 5.0). Inter-rater agreement was high, with a Cohen's κ of 0.84 (95% CI: 0.76–0.92), indicating strong rating consistency across independent experts.

Error Analysis. We subsequently analyzed the 7 cases where TSP annotations received an AS < 4 and observed three failure patterns: 4 of 7 involved Rare tumor types with extremely low contrast between the tumor mass and surrounding edema, resulting in highly ambiguous boundaries; 2 of 7 corresponded to very small lesions (volume fraction $< 0.05\%$, maximum cross-sectional area < 25 mm^2), where localization accuracy was limited by inherent imaging resolution; the remaining case presented multifocal enhancing patterns, which fall outside the single-lesion assumption underlying the TSP pipeline. These failure modes also reflect the intrinsic diagnostic difficulty of deep-seated tumor analysis and point to directions for future improvement.

Overall, these results demonstrate that TSP provides stable and clinically acceptable spatial priors in most cases, while limitations remain in scenarios involving indistinct boundaries, extremely small lesions, and complex lesion distributions. These challenging cases also reflect the intrinsic difficulty of the DeepSite-MRI task and highlight important directions for future improvements.

For additional context, we also evaluate existing open-source medical image segmentation models, focusing on MedSAM3 as a representative example. Specifically, we evaluate it under the same fully automated setting, without manual spatial prompts or case-specific correction. In this setting, MedSAM3 often failed to consistently localize deep-seated tumors, whereas TSP generated usable tumor-centered spatial priors. A detailed comparison under prompt-free, text-prompted, and expert-assisted settings is provided in Appendix B.

4 Benchmark Experiments

This section establishes reproducible baselines for the community. We describe the experimental setup (Sec. 4.1), report classification results across 10 architectures under three prior conditions (Sec. 4.2), and examine the data efficiency of existing architectures to reveal their limitations and open research opportunities on this task (Sec. 4.3).

4.1 Experimental Setup

Experimental Conditions. All baselines use single-modality T1-CE as input to eliminate confounds from multi-modal fusion. We evaluate under three conditions: (i) *Without Prior*: no spatial information, serving as a pure vision baseline; (ii) *Coarse Prior*: only the first-stage TSP output (VLM bounding boxes + Gaussian smoothing); and (iii) *Refined Prior*: the full two-stage TSP mask. These progressive conditions enable the main experiment to inherently function as an ablation study. The spatial prior is integrated via early fusion: the prior mask is concatenated with the T1-CE volume as an additional input channel, requiring no backbone modifications.

Metrics. We report accuracy, precision, recall, and F1-score. Precision measures the ability to avoid false positives; recall captures sensitivity to true positives. The F1-score balances the two, providing a more robust measure under class imbalance.

Implementation. To ensure the robustness and reliability of our evaluation, all experiments are conducted using a 5-fold cross-validation strategy. Our method is implemented in PyTorch and trained on NVIDIA RTX 4090 GPUs for up to 250 epochs using the AdamW optimizer with a learning rate of 5×10^{-5} and a weight decay of 1×10^{-4}. The learning rate was scheduled using cosine annealing

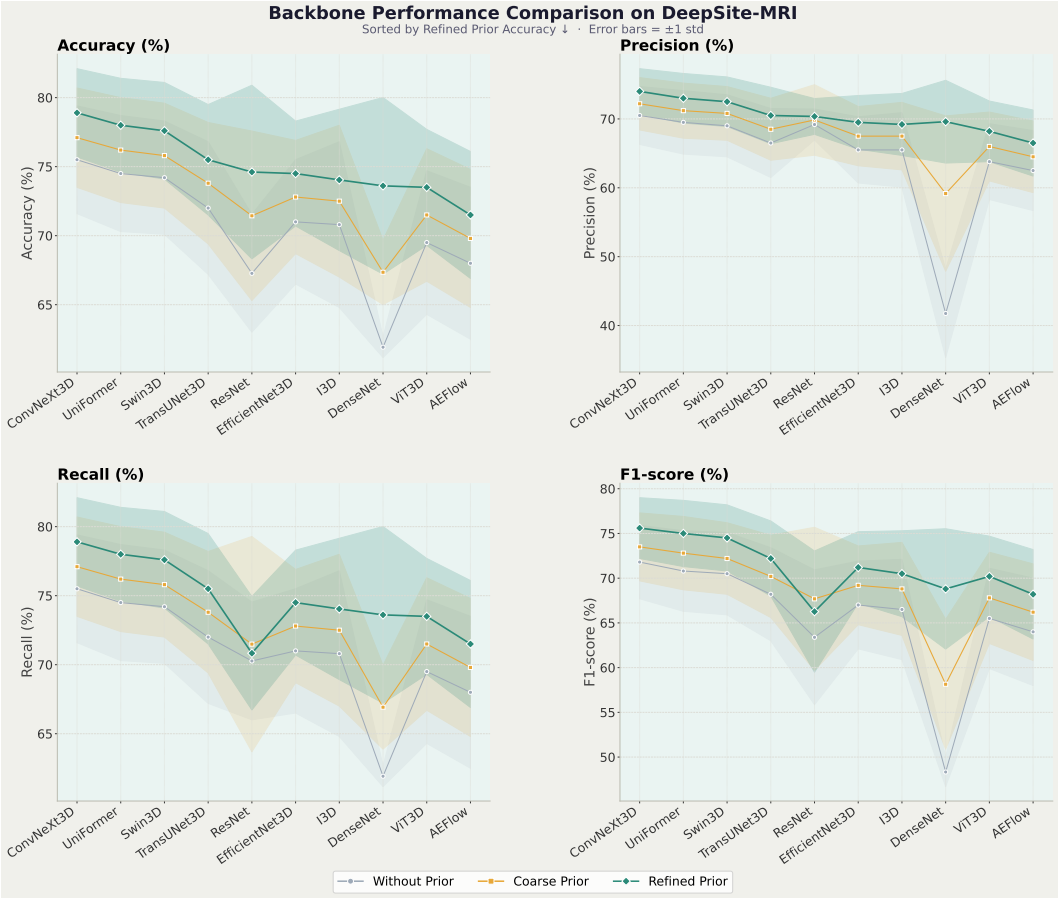


Figure 3: Classification accuracy across all 10 backbones under three prior conditions (Without Prior / Coarse Prior / Refined Prior). Shaded regions denote ± 1 std over 5-fold cross-validation. Full numerical results are reported in **Appendix E**, Tab. 4.

with warm restarts (initial period $T_0 = 10$ epochs). Gradient clipping (max norm 2.0) and dropout ($p = 0.2$) were applied for stability. Label smoothing ($\epsilon = 0.05$) was used to reduce overconfident predictions. To simulate an effective batch size of 4 while preserving memory efficiency, training adopted gradient accumulation over two steps with a batch size of 2.

4.2 Experimental Results

To assess whether the proposed spatial prior generalizes across architectures of different types, we evaluate 10 backbones spanning CNNs (ResNet3D He et al. (2016), DenseNet3D Huang et al. (2017), ConvNeXt3D Woo et al. (2023), EfficientNet3D Tan and Le (2019), I3D Carreira and Zisserman (2017)) and Transformer-based models (ViT3D Dosovitskiy et al. (2020), Swin3D Liu et al. (2021), UniFormer Li et al. (2023), TransUNet3D Chen et al. (2021), AEFlow Zhao et al. (2023)), all initialized with ImageNet or Kinetics pre-trained weights and adapted to a 4-class classification head. As shown in Fig. 3, we evaluate all backbones under the three prior conditions; full numerical results are provided in **Appendix E**.

The results show consistent benefits from spatial priors across architectures. Refined Prior improves accuracy over Without Prior for all 10 backbones, with gains of 2.50–11.69% (average +4.76%). The additional +2.44% gain from Coarse to Refined Prior indicates a complementary contribution from the MLP refinement stage. ConvNeXt3D and UniFormer achieve the highest accuracies of 78.90% and 78.00%.

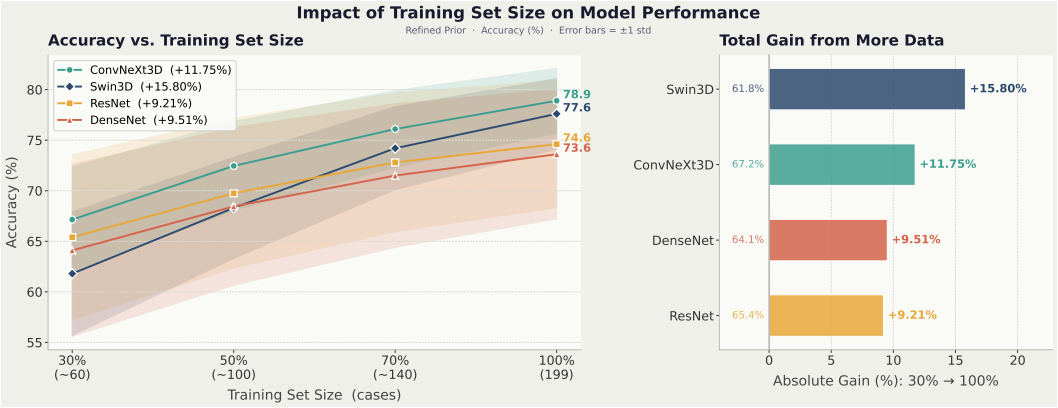


Figure 4: Data efficiency curves across four representative backbones under Refined Prior. Shaded regions denote ±1 std over 5-fold cross-validation, highlighting the high variance at low data fractions. Full results are shown in **Appendix F**, Tab. 6.

Despite these gains, the task remains challenging. The best accuracy remains substantially below that reported for similar architectures on larger brain tumor benchmarks, reflecting the difficulty of small, visually ambiguous deep-seated lesions and the severe class imbalance, with only 10 Rare cases (4.0%). Given that the same architectures typically achieve stronger performance on larger-scale datasets, a key question arises: are existing architectures data-efficient enough for such clinically challenging tasks?

4.3 Data Efficiency Analysis

To investigate this question, we evaluate four representative backbones under the Refined Prior setting across training set sizes of 30%, 50%, 70%, and 100%, with the test set held fixed. Results are shown in Fig. 4; complete numerical results are provided in **Appendix F**, Tab. 6.

The scaling analysis reveals substantial data-efficiency limitations. None of the four representative backbones approaches its full-data performance with only 30% of the training data, and all models exhibit higher fold-wise variance in low-data regimes. Swin3D is particularly sensitive, improving by +15.80 percentage points from 30% to 100% training data, compared with +11.75% for ConvNeXt3D and approximately +9% for ResNet and DenseNet. These results highlight the difficulty of learning robust representations from limited data and establish DeepSite-MRI as a useful testbed for data-efficient learning.

5 Benchmark Value and Open Research Directions

Beyond classification, the biopsy-confirmed labels, multi-level spatial annotations, and realistic long-tail distribution of DeepSite-MRI open several research directions (detailed in **Appendix H**).

Transferability of the annotation methodology. The annotation paradigm of the TSP pipeline (zero-shot VLM coarse localization + MLP refinement + expert review) offers a reusable blueprint for data-scarce medical imaging domains. When conventional semi-automatic tools cannot be initialized due to the lack of in-domain pretrained models, leveraging general-purpose VLMs for zero-shot localization, followed by lightweight refinement to approach expert-level quality, can compress per-case annotation cost from tens of minutes to a few minutes. This paradigm is transferable to dataset construction for other rare pathologies.

Learning under data-limited regimes. The dataset exhibits a clinically realistic and irreducible long-tail distribution (the *Rare* category contains only 10 instances, 4.0%), an imbalance arising from the epidemiological characteristics of the disease rather than sampling bias, providing a clinically grounded benchmark for long-tail recognition and few-shot adaptation. Meanwhile, the combination of biopsy-confirmed labels and multi-level spatial annotations naturally supports weakly supervised segmentation and data-efficient transfer learning. Our data scaling experiments have demonstrated

that no architecture has reached saturation at the current scale, further validating the need for efficient learning strategies in small-data regimes.

Clinical translational directions. From a clinical perspective, the dataset directly supports multiple translational research directions: radiomics biomarker mining tailored to specific MRI phenotypic subtypes, AI-assisted biopsy target planning, and surgical risk stratification based on deep anatomical locations. These directions collectively address the core clinical motivation of this work: reducing reliance on invasive biopsies through reliable preoperative MRI analysis.

6 DISCUSSION

Stereotactic biopsy for deep intracranial tumors is an inherently low-frequency procedure, making data scarcity a fundamental constraint rather than an oversight. Within this constraint, we have reduced the spatial annotation cost to near zero through the open-source TSP pipeline, enabling the construction of the largest public benchmark (249 cases) for this pathology to date. **This work provides a starting point, not an end point.** The open-source annotation capability of TSP can empower researchers worldwide to generate spatial annotations on their own biopsy cohorts at minimal cost, thereby breaking through the data boundary of any single institution. On our side, we are already actively pursuing further multi-center collaboration with a concrete plan to expand the dataset to **1,000+ cases** covering diverse demographics and scanning protocols. With continued data expansion from both our team and the broader research community, we expect this benchmark to support increasingly impactful research in the field. For example, once sufficient volume and demographic diversity are achieved, the dataset could serve as a foundation for pre-training domain-specific foundation models for deep intracranial tumor analysis.

7 CONCLUSION

This paper presents DeepSite-MRI, the first public MRI benchmark for deep intracranial tumors, comprising 249 biopsy-confirmed cases with multi-level spatial annotations, together with the open-source TSP pipeline that generates spatial priors at near-zero manual cost. We systematically evaluate 10 mainstream architectures spanning CNNs and Transformers on this benchmark; even well-established methods that excel on large-scale brain tumor benchmarks underperform here, and all tested architectures exhibit fundamental data efficiency deficiencies. These findings highlight the unique challenges of deep intracranial tumor diagnosis and establish DeepSite-MRI as a valuable testbed for data-constrained learning research, providing a solid foundation for future work in long-tailed recognition, data-efficient transfer learning, and clinical translation.

AI USE STATEMENT

Generative AI tools were used during the preparation of this manuscript for language editing, improving clarity and coherence, LaTeX formatting, literature search and summarization, and assistance with code debugging and implementation-related issues. AI-generated suggestions were used only as supplementary assistance and were reviewed, revised, and verified by the authors before being incorporated into the manuscript or research workflow.

The research questions, dataset construction, annotation methodology, experimental design, evaluation protocol, interpretation of experimental results, and scientific conclusions were determined by the authors. Generative AI tools were not used to generate or alter experimental data, fabricate experimental results, or determine the scientific conclusions reported in this work. AI-assisted code and implementation suggestions were tested against the actual experimental pipeline, and AI-assisted factual claims and references were checked against the corresponding original sources.

The authors take full responsibility for the final content of this work, including all scientific claims, experimental results, figures, tables, code, and conclusions.

REFERENCES

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13:4128, 2022. doi: 10.1038/s41467-022-30695-9.

Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). *Insight j*, 2 (365):1–35, 2009.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *The Cancer Imaging Archive*, 2017. doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Jun Cheng. Brain tumor dataset. *figshare*, 2017. doi: 10.6084/m9.figshare.1512427.v5.

Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Rayan Fawaz, Mehdi Bensemain, Karima Mokhtari, Lucia Nichelli, and Bertrand Mathon. Evaluating hemorrhagic risks using intratumoral susceptibility signals in stereotactic brain biopsies. *Neurosurgical Review*, 48(1):451, 2025.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Guanghui Fu, Lucia Nichelli, Darío Herrán de la Gala, Sophie Loizillon, Camile Bousfiha, Romain Valabregue, Agusti Alentorn, Khê Hoang-Xuan, Bertrand Mathon, Carole Soussain, et al. Automatic segmentation of primary central nervous system lymphoma at clinical routine postcontrast t1-weighted mri. *Radiology: Imaging Cancer*, 7(5):e240446, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Mathias Holtkamp, Vicky Parmar, René Hosch, Luca Salhöfer, Hanna Styczen, Yan Li, Marcel Opitz, Martin Glas, Nika Guberina, Karsten Wrede, et al. Ai-guided virtual biopsy: Automated differentiation of cerebral gliomas from other benign and malignant mri findings using deep learning. *Neuro-Oncology Advances*, 7(1):vdae225, 2025.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17): 4952–4964, 2019.

Anahita Fathi Kazerooni, Nastaran Khalili, Ariana M. Familiar, et al. The brain tumor segmentation (BraTS) challenge 2023: Focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs), 2023a.

Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Ariana M. Familiar, Jenna Albrecht, Maruf Adewole, Syed Muhammad Anwar, et al. A multi-institutional pediatric dataset of clinical radiology MRIs by the Children's Brain Tumor Network (CBTN). *Scientific Data*, 10:845, 2023b. doi: 10.1038/s41597-023-02881-x.

Dominic LaBella, Katherine Schumacher, Michael Mix, Kevin Leu, Shan McBurney-Lin, Pierre Nedelec, Javier Villanueva-Meyer, David R Raleigh, Jonathan Shapey, Tom Vercauteren, et al. The 2024 brain tumor segmentation challenge meningioma radiotherapy (brats-men-rt) dataset. *Scientific Data*, 2026.

Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12581–12600, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yann Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10): 1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

Ahmed W. Moawad, Rachit Jain, Rachit Bhalerao, et al. The brain tumor segmentation (BraTS-METS) challenge 2023: Brain metastasis segmentation on pre-treatment MRI, 2023.

Masoud Nickparvar. Brain tumor MRI dataset. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset, 2021.

Msoud Nickparvar. Brain tumor MRI dataset, 2024.

Mackenzie Price, Christine Ann Pittman Ballard, Julia R Benedetti, Carol Kruchko, Jill S Barnholtz-Sloan, and Quinn T Ostrom. Cbtrus statistical report: Pediatric brain tumor foundation childhood and adolescent primary brain and other central nervous system tumors diagnosed in the united states in 2017-2021. *Neuro-Oncology*, 27(Supplement_1):i1–i42, 2025.

Qwen Team. Qwen3-VL: A vision-language model with high-resolution visual understanding. https://github.com/QwenLM/Qwen3-VL, 2025. Accessed: 2025-01-08.

Jeffrey D. Rudie, Rachit Saluja, David A. Weiss, Timothy N. Watkins, Jianhua Yao, Leo P. Sugrue, and Javier E. Villanueva-Meyer. The University of California San Francisco brain metastases stereotactic radiosurgery (UCSF-BMSR) MRI dataset. *Radiology: Artificial Intelligence*, 6(2): e230126, 2024. doi: 10.1148/ryai.230126.

Toni Saad, Alexander Tuck, Farhad Golestani, Paul Smith, and Rory McCulloch. Primary central nervous system lymphoma: a practical guide for neurologists. *Practical Neurology*, 23(4):286–292, 2023.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128(2):336–359, 2020.

Suchada Supbumrung, Anukoon Kaewborisutsakul, and Thara Tunthanathip. Machine learning-based classification of pineal germinoma from magnetic resonance imaging. *World Neurosurgery: X*, 20: 100231, 2023.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023.

Hao Zhang, Ruizhe Wang, Yuanqiang Yu, Jinfang Liu, Tianmeng Luo, and Fan Fan. Glioblastoma treatment modalities besides surgery. *Journal of Cancer*, 10(20):4793, 2019.

Yuzhong Zhao, Qiaoqiao Ding, and Xiaoqun Zhang. Ae-flow: Autoencoders with normalizing flows for medical images anomaly detection. In *The Eleventh International Conference on Learning Representations*, 2023.

APPENDIX

A TSP: COMPLETE IMPLEMENTATION DETAILS

A.1 BLOCK 1: VLM-GUIDED SPATIAL COARSE-GRAINED LOCALIZATION

Due to the lack of voxel-level annotations in the collected data, we leverage the powerful multimodal reasoning capabilities of a Vision-Language Model (specifically, Qwen3-VL-Thinking (Qwen Team, 2025)) to perform coarse-grained spatial localization of deep-seated intracranial tumors. *The core objective of this block is to extract clinically meaningful spatial priors from 3D MRI volumes without any manual annotation, providing global guidance for the fine-grained refinement in Block 2.*

Step 1: Standardized MRI Pre-processing. Raw 3D MRI data typically exhibit large variability across scanners and subjects, including systematic differences in spatial resolution, contrast, and signal-to-noise ratio; feeding such data directly into VLM inference would substantially degrade cross-subject localization consistency. To this end, we apply a standardized three-stage pre-processing pipeline uniformly to all T1-CE images: (i) skull stripping using HD-BET (Isensee et al., 2019) to remove non-brain tissue; (ii) bias field correction with ANTs N4 (Avants et al., 2009) to mitigate intensity inhomogeneity; and (iii) nonlinear registration via ANTs SyN (Avants et al., 2009) to align individual scans to a standardized anatomical template (MNI space). Let the raw MRI volume be denoted as $\mathbf{X}_{\text{raw}} \in \mathbb{R}^{D\times H\times W}$. The pre-processing pipeline can be formulated as:

$$\mathbf{X}_{\text{proc}} = \mathcal{F}_{\text{SyN}}(\mathcal{F}_{\text{N4}}(\mathcal{F}_{\text{BET}}(\mathbf{X}_{\text{raw}}))) \,, \tag{3}$$

where $\mathbf{X}_{\text{proc}} \in \mathbb{R}^{D\times H\times W}$ denotes the standardized processed MRI volume; $\mathcal{F}_{\text{BET}}$, $\mathcal{F}_{\text{N4}}$, and $\mathcal{F}_{\text{SyN}}$ represent the operations of brain extraction, bias field correction, and nonlinear spatial normalization, respectively; and D, H, W correspond to the spatial dimensions (depth, height, and width). The resulting $\mathbf{X}_{\text{proc}}$ is then passed to the subsequent VLM inference step.

Step 2: Structured Radiological Prompt-guided Bounding Box Prediction. Since the VLM cannot directly process 3D volumetric data, we perform axial slicing on $\mathbf{X}_{\text{proc}}$ to convert it into a sequence of 2D cross-sectional images:

$$\left\{\mathbf{X}_{\text{slice}}^{(d)}\right\}_{d=1}^{D} = \mathcal{F}_{\text{3D}\to\text{2D}}(\mathbf{X}_{\text{proc}}) \,, \tag{4}$$

where $\mathcal{F}_{\text{3D}\to\text{2D}}$ denotes the 3D-to-2D axial slicing operation, and $\mathbf{X}_{\text{slice}}^{(d)} \in \mathbb{R}^{H\times W}$ is the d-th axial slice. Subsequently, we construct a Structured Radiological Prompt $\mathcal{P}_{\text{radio}}$ for each slice, directing the model to assess tumor likelihood based on key diagnostic cues (including abnormal signal intensity patterns, irregular enhancement boundaries, edema-associated gradients, and anatomical asymmetry relative to the midline) and to generate predicted bounding boxes:

$$\mathcal{B}^{(d)} = \mathcal{F}_{\text{VLM}}\left(\mathbf{X}_{\text{slice}}^{(d)},\ \mathcal{P}_{\text{radio}}\right) , \tag{5}$$

where $\mathcal{F}_{\text{VLM}}$ denotes the inference operation of the Qwen3-VL-Thinking model; $\mathcal{B}^{(d)} = \left(x_{\min}^{(d)},\ y_{\min}^{(d)},\ x_{\max}^{(d)},\ y_{\max}^{(d)}\right)$ are the predicted bounding box coordinates on the d-th slice, and $\mathcal{B}^{(d)} = \emptyset$ if no credible lesion evidence is identified on that slice.

Step 3: 3D Reconstruction and Spatial Continuity. We spatially stack all slices containing valid predicted bounding boxes to preliminarily reconstruct a 3D candidate region. However, the 3D volume generated by directly stacking 2D bounding boxes exhibits spatial discreteness and discontinuity, manifesting as stair-step artifacts; furthermore, per-slice prediction biases of the VLM may result in incomplete tumor coverage. To address both issues, we sequentially apply Gaussian smoothing ($\sigma = 1.5$ voxels) and morphological dilation (spherical structuring element, radius $= 3$ voxels), transforming the discrete bounding box sequence into a continuous probability distribution and yielding a smooth 3D spatial prior map that maximally covers the full tumor extent:

$$\mathbf{X}_{\text{tumor-coarse}} = \mathcal{F}_{\text{smooth}}\left(\mathcal{F}_{\text{2D}\to\text{3D}}\left(\left\{\mathcal{B}^{(d)}\right\}_{d=1}^{D},\ \left\{\mathbf{X}_{\text{slice}}^{(d)}\right\}_{d=1}^{D}\right)\right) , \tag{6}$$

where $\mathcal{F}_{\text{2D}\to\text{3D}}$ denotes the 3D reconstruction operation that activates valid bounding box regions as binary voxel masks and stacks them along the z-axis; $\mathcal{F}_{\text{smooth}}$ denotes the sequential application of Gaussian smoothing and morphological dilation; and $\mathbf{X}_{\text{tumor-coarse}} \in \mathbb{R}^{D\times H\times W}$ is the resulting coarse-grained 3D tumor spatial prior map. *This map serves as an initial coarse-grained global attention, providing spatial guidance for the fine-grained localization refinement in Block 2 (Eq. 7).*

A.2 Block 2: Feature-driven Fine-grained Localization

Leveraging the coarse-grained tumor spatial prior $\mathbf{X}_{\text{tumor-coarse}}$ generated in Block 1 (Eq. 6), we obtain weakly supervised voxel-level labels and initiate a self-correction process. *This step is necessary because probability maps derived directly from bounding boxes are spatially over-smoothed: they fail to conform to irregular tumor boundaries and inevitably incorporate surrounding non-tumor background tissue as candidate regions.* For deep-seated intracranial tumors, this limitation is particularly pronounced: the signal gradient between the tumor core and the surrounding edema band is often indistinct, causing VLM-predicted bounding boxes to systematically overestimate the lesion extent and introduce abundant background voxels that degrade the quality of discriminative features extracted by the downstream classifier.

To address this, we treat the output of Block 1 as pseudo-labels and select high-confidence voxels (those whose coarse prior value exceeds threshold τ) as positive training samples to train a lightweight voxel-wise Multi-Layer Perceptron (MLP) classifier $\mathcal{F}_{\text{MLP}}$. The MLP takes as input a feature vector concatenating the local MRI signal intensity, spatial gradient magnitude, local texture statistics, and the VLM prior value at each voxel, enabling it to learn the intrinsic visual discriminability between tumor and background tissue in terms of MRI signal and texture patterns.

We then take the *union* of the MLP's refined prediction and the VLM-guided coarse prior (Union Validation), so as to maximize the recall of potential tumor voxels and compensate for any under-segmentation produced by the MLP when trained on imperfect pseudo-labels. The full process can be formulated as:

$$\mathbf{X}_{\text{tumor-refined}} = \mathbf{X}_{\text{tumor-coarse}} \cup \mathcal{F}_{\text{MLP}}(\mathbf{X}_{\text{img}} \mid \mathbf{X}_{\text{tumor-coarse}} > \tau), \tag{7}$$

where $\mathcal{F}_{\text{MLP}}$ denotes the voxel-level prediction of the MLP trained on high-confidence regions (thresholded by τ); $\cup$ represents the voxel-wise union operation; $\mid$ denotes a conditional indexing operation, indicating that the MLP is trained exclusively on voxels where the VLM-based coarse prior confidence exceeds τ; and $\mathbf{X}_{\text{tumor-refined}} \in \mathbb{R}^{D\times H\times W}$ is the resulting refined spatial prior map, which is subsequently concatenated with the original T1-CE volume as an additional input channel to the downstream classification backbone.

Unlike the coarse global guidance produced in Block 1, the refinement in Block 2 is locally adaptive: the MLP is trained independently for each patient based on that individual's MRI signal distribution. This design allows the refined prior to preserve the global localization advantage of the VLM while effectively correcting boundary deviations and suppressing background noise, achieving a complementary fusion of coarse-grained global prior and fine-grained local feature consistency.

B Comparison with MedSAM3

To further contextualize the spatial prior generated by TSP, we compared it with MedSAM3 under the same fully automated annotation setting. Our objective was not to establish a general segmentation ranking, but to examine whether a general-purpose medical segmentation model can provide reliable localization of deep-seated tumors without case-specific manual spatial prompts or correction.

We first evaluated MedSAM3 without any spatial prompt or post-processing. Under this setting, MedSAM3 frequently responded to the brain parenchyma or brain contour rather than consistently localizing the deep-seated tumors. This behavior highlights a prompt-free cold-start difficulty for our dataset: although general-purpose medical segmentation models provide broad concept segmentation capabilities, they do not necessarily provide reliable tumor localization when no spatial cue is available.

For a more controlled comparison, we additionally evaluated a fixed text-prompted setting using the same "brain tumor" prompt for all cases, without case-specific prompt optimization. Following expert delineation of tumor cores on an independent set of 50 cases, we used the expert annotations as the reference standard and compared the following settings:

The comparison shows that, without manual assistance, TSP achieves higher spatial overlap and lower boundary distance than the fixed text-prompted MedSAM3 setting. MedSAM3 can achieve higher segmentation accuracy when expert prompts and case-specific corrections are provided; however,

Table 2: **Comparison of automated spatial localization settings against expert-delineated tumor cores on an independent set of 50 cases.** Dice and IoU are higher-is-better, whereas HD95 is lower-is-better. "Loc. Fail." denotes cases in which the method failed to provide a usable tumor localization under the corresponding setting.

Method	Manual	Dice ↑	IoU ↑	HD95 (mm) ↓	Loc. Fail.
MedSAM3, without any prompt	No	0.02 ± 0.02	0.01 ± 0.01	66.4 ± 14.8	50/50
MedSAM3, fixed "brain tumor" text prompt	No	0.56 ± 0.22	0.42 ± 0.21	18.6 ± 16.3	6/50
TSP tumor-only candidate mask	No	0.72 ± 0.11	0.57 ± 0.13	7.6 ± 5.4	2/50
MedSAM3 + expert-assisted correction	Yes	0.80 ± 0.10	0.68 ± 0.12	5.2 ± 3.8	0/50

Table 3: **Definitions of the two independent expert evaluation metrics used in the double-blind validation study.** Accuracy Score (AS) is a five-point Likert scale that assesses the spatial localization quality of TSP-generated priors, where AS ≥ 4 is considered clinically acceptable in terms of boundary coverage. Clinical Acceptability (CA) is a binary criterion indicating whether an annotation can serve as a research-grade reference without further manual correction.

Metric	Score	Definition
Accuracy Score (AS)	5	Complete tumor coverage with well-defined boundaries and no evident over-segmentation.
	4	Largely complete coverage with minor boundary deviations.
	3	Main region covered but with noticeable under- or over-segmentation.
	2	Substantial localization error.
	1	Failure to localize or grossly incorrect annotation.
Clinical Acceptability (CA)	Yes	Annotation can serve as a reliable research-grade reference without further manual correction.
	No	Annotation requires additional manual refinement before use.

this setting requires manual intervention and therefore addresses a different problem from the fully automated annotation pipeline considered by TSP. Accordingly, we do not interpret these results as demonstrating that TSP is superior to MedSAM3 for interactive segmentation. Rather, the comparison illustrates the distinction between interactive segmentation performance and prompt-free, scalable generation of clinically usable spatial priors.

It is important to distinguish the tumor-only candidate mask used for segmentation evaluation from the complete Refined Prior used for downstream classification. The latter intentionally retains some peritumoral edema and relevant surrounding imaging context, which may contain diagnostically useful information. In contrast, expert delineations typically focus on the tumor core. Therefore, the Dice, IoU, and HD95 values reported above are calculated exclusively using the tumor-only masks separately exported by TSP. The complete spatial priors are instead evaluated through expert-rated clinical acceptability, lesion coverage, and downstream classification utility.

C Detailed Display Rubrics in Evaluation Setup

This appendix details the complete assessment criteria (AC) established for the independent expert evaluation of our method. To rigorously and objectively measure the performance of the generated annotations, we employed two primary metrics during the double-blind validation study: the Accuracy Score (AS) and Clinical Acceptability (CA). The AS evaluates the spatial localization quality and boundary adherence of the generated priors on a five-point Likert scale, while CA provides a binary assessment of whether the annotation is of sufficient quality to serve as a research-grade reference without further human intervention. The comprehensive definitions and scoring rubrics for both metrics are provided in Tab. 3.

D Complete Benchmark Results

Tab. 4 reports the complete evaluation results of all 10 backbones under three spatial prior conditions (Without Prior, Coarse Prior, Refined Prior), across all four metrics (Accuracy, Precision, Recall, F1-score). Results are reported as mean ± standard deviation over 5-fold cross-validation and are sorted by accuracy under Refined Prior in descending order.

E Full Main Experimental Results

To comprehensively assess the effectiveness of our proposed pipeline, we evaluate 10 representative architectures prevalent in contemporary 3D medical image analysis and video understanding, including Convolutional Neural Networks (ResNet3D, DenseNet3D, ConvNeXt3D, EfficientNet3D, I3D) and Transformer-based models (ViT3D, Swin3D, UniFormer, TransUNet3D, AEFlow). All models are initialized with ImageNet or Kinetics pre-trained weights, and their classification heads are replaced with a 4-class linear layer. Complete numerical results across all three prior conditions and all four metrics (Accuracy, Precision, Recall, F1-score) are provided in Tab. 4.

Table 4: **Classification performance of 10 mainstream 3D architectures on DeepSite-MRI under three spatial prior conditions.** The evaluated settings include No Prior (pure vision baseline), Coarse (coarse vision-language model localization only, corresponding to TSP Stage 1), and Refined (full two-stage TSP pipeline). Results are reported as mean ± standard deviation over 5-fold cross-validation and are sorted by accuracy under Refined Prior in descending order. Acc, Prec, Rec, and F1 denote accuracy, precision, recall, and F1-score, respectively. All values are percentages.

Backbone	Condition	Acc	Prec	Rec	F1
ConvNeXt3D Woo et al. (2023)	No Prior	75.50±3.90	70.50±4.20	75.50±3.90	71.80±4.10
	Coarse	77.10±3.60	72.20±3.80	77.10±3.60	73.50±3.80
	Refined	78.90±3.20	74.00±3.30	78.90±3.20	75.60±3.40
UniFormer Li et al. (2023)	No Prior	74.50±4.20	69.50±4.60	74.50±4.20	70.80±4.50
	Coarse	76.20±3.80	71.20±4.00	76.20±3.80	72.80±4.10
	Refined	78.00±3.40	73.00±3.60	78.00±3.40	75.00±3.70
Swin3D Liu et al. (2021)	No Prior	74.20±4.10	69.00±4.50	74.20±4.10	70.50±4.60
	Coarse	75.80±3.80	70.80±3.90	75.80±3.80	72.20±4.00
	Refined	77.60±3.50	72.50±3.60	77.60±3.50	74.50±3.70
TransUNet3D Chen et al. (2021)	No Prior	72.00±4.80	66.50±5.00	72.00±4.80	68.20±5.20
	Coarse	73.80±4.40	68.50±4.50	73.80±4.40	70.20±4.60
	Refined	75.50±4.00	70.50±4.10	75.50±4.00	72.20±4.20
ResNet He et al. (2016)	No Prior	67.27±4.26	69.19±2.24	70.27±4.26	63.39±7.53
	Coarse	71.44±6.14	69.85±5.12	71.49±7.80	67.69±8.00
	Refined	74.61±6.29	70.36±2.60	70.83±4.11	66.26±6.78
EfficientNet3D Tan and Le (2019)	No Prior	71.00±4.50	65.50±4.80	71.00±4.50	67.00±4.90
	Coarse	72.80±4.10	67.50±4.30	72.80±4.10	69.20±4.40
	Refined	74.50±3.80	69.50±3.90	74.50±3.80	71.20±4.00
I3D Carreira and Zisserman (2017)	No Prior	70.80±6.00	65.50±5.40	70.80±6.00	66.50±5.60
	Coarse	72.50±5.50	67.50±4.90	72.50±5.50	68.80±5.20
	Refined	74.04±5.12	69.20±4.50	74.04±5.12	70.50±4.80
DenseNet Huang et al. (2017)	No Prior	61.92±0.75	41.75±6.40	61.92±0.75	48.33±1.63
	Coarse	67.35±2.34	59.16±11.28	66.92±3.04	58.14±7.26
	Refined	73.61±6.40	69.60±6.01	73.61±6.40	68.80±6.73
ViT3D Dosovitskiy et al. (2020)	No Prior	69.50±5.20	63.80±5.50	69.50±5.20	65.50±5.60
	Coarse	71.50±4.80	66.00±5.00	71.50±4.80	67.80±5.10
	Refined	73.50±4.20	68.20±4.40	73.50±4.20	70.20±4.50
AEFlow Zhao et al. (2023)	No Prior	68.00±5.50	62.50±5.80	68.00±5.50	64.00±6.00
	Coarse	69.80±5.00	64.50±5.20	69.80±5.00	66.20±5.40
	Refined	71.50±4.60	66.50±4.80	71.50±4.60	68.20±5.00

Table 5: **Class-wise recall of representative models on DeepSite-MRI.** Results are reported as mean ± standard deviation across the five patient-level cross-validation folds. Acc. and Rec. denote accuracy and recall, respectively. All values are percentages.

Model	Overall Acc.	Glioma Rec.	PCNSL Rec.	Germinoma Rec.	Rare Rec.
ConvNeXt3D	78.90 ± 3.20	91.3 ± 1.8	69.3 ± 6.6	38.3 ± 11.2	30.0 ± 27.4
UniFormer	78.00 ± 3.40	92.0 ± 1.8	68.0 ± 6.5	33.3 ± 10.2	20.0 ± 27.4
ResNet3D	74.61 ± 6.29	90.6 ± 4.3	59.5 ± 8.6	35.0 ± 18.1	20.0 ± 27.4

Class-wise Recall Analysis. We further report class-wise recall for three representative models: ConvNeXt3D, UniFormer, and ResNet3D.

These results show that overall accuracy is primarily driven by performance on Glioma and PCNSL, whereas Germinoma and Rare remain substantially more difficult. This pattern is consistent across the representative CNN and Transformer models. The limited sample sizes of the minority classes also lead to pronounced fold-wise variability. Germinoma contains only 18 cases, corresponding to approximately 3–4 cases per validation fold; consequently, a single case can change the recall of one fold by approximately 25–33%. Rare contains only 10 cases, with two cases per fold, so a single case can produce a 50% change in fold-wise recall.

The Rare class additionally comprises several low-frequency pathologies, resulting in substantial intra-class heterogeneity. These factors make stable representation learning particularly difficult for the minority classes. The reported results should therefore be interpreted as exploratory baselines for minority-class recognition rather than evidence of reliable clinical diagnostic capability.

We retain these cases to faithfully represent the long-tailed distribution encountered in deep intracranial tumor diagnosis and to provide a public evaluation starting point for few-shot learning, class-imbalanced learning, and rare-pathology recognition. DeepSite-MRI therefore also provides a realistic benchmark for studying learning under severe class imbalance and limited samples. **Consistent Gains from Spatial Priors.** The incorporation of spatial priors yields consistent and significant performance improvements. Refined Prior outperforms the Without Prior baseline across all 10 backbones and all evaluation metrics, with ACC gains ranging from +2.50% (ResNet) to +11.69% (DenseNet), averaging +4.76%. This cross-architecture consistency indicates that the lesion localization cues encoded in the TSP-derived spatial priors are broadly effective, and their benefits are not confined to specific receptive field designs or attention mechanisms. Notably, the further improvement from Coarse Prior to Refined Prior (averaging +2.44%) suggests that the MLP-based refinement stage provides additional gains beyond the coarse localization from the VLM, demonstrating that both stages contribute independently to the final performance. Among all models, ConvNeXt3D and UniFormer achieve the best results under the Refined Prior setting, reaching 78.90% and 78.00%, respectively, establishing new state-of-the-art baselines on DeepSite-MRI. In contrast, ViT3D (73.50%) and AEFlow (71.50%) rank relatively lower, with AEFlow exhibiting the weakest starting point at only 68.00% under the Without Prior setting. Nevertheless, all models benefit from the introduction of Refined Prior, with even the lowest-performing AEFlow achieving an improvement of +3.50 percentage points.

Task Difficulty of the Dataset and Extreme Class Imbalance. Despite the improvements, the overall task remains highly challenging. Even the best-performing model, ConvNeXt3D with Refined Prior, achieves an ACC of only 78.90%, whereas similar architectures typically exceed 90% on larger-scale brain tumor datasets (e.g., BraTS2021, binary classification with over 1,000 samples). This performance gap is largely attributable to the inherent difficulty of deep-seated small lesions, the visual ambiguity of the target pathologies, and notably, extreme class imbalance. Specifically, the *Rare* category comprises only 10 instances (4.0% of the entire dataset). Under our 5-fold cross-validation protocol, this results in merely 2 *Rare* cases per validation fold. This severe data scarcity severely restricts the models' ability to learn robust representations for this specific category, leading to frequent misclassifications that disproportionately penalize the overall performance metrics. Furthermore, this extreme minority class introduces significant statistical instability during evaluation; misclassifying just one or two rare cases causes dramatic fluctuations in fold-wise results. This instability is reflected in the 5-fold standard deviations, which generally fall between 3% and 6%. Therefore, the current performance ceiling is strongly constrained by the long-tailed nature of the

dataset, indicating that DeepSite-MRI continues to offer strong discriminative power and intrinsic challenges for benchmarking future methodological advances.

F Data Efficiency Analysis

All the above results are based on the full dataset of 249 cases. To assess the data efficiency of existing architectures on this clinically challenging task, we conduct a systematic study under the Refined Prior setting, while keeping the test set fixed. Specifically, we evaluate four representative backbones across training set sizes of 30% / 50% / 70% / 100%, aiming to reveal whether current architectures can learn effective representations from limited data or whether they exhibit fundamental efficiency deficiencies. Complete numerical results are provided in Tab. 6.

Table 6: **Accuracy (%, mean ± standard deviation over 5-fold cross-validation) of four representative backbones as a function of training set size.** Training set sizes correspond to 30%, 50%, 70%, and 100% of the training cases (approximately 60, 100, 140, and 199 cases, respectively). All models are evaluated under the Refined Prior condition with a fixed test set.

Backbone	**30% (∼60)**	**50% (∼100)**	**70% (∼140)**	**100% (199)**
ConvNeXt3D Woo et al. (2023)	67.15±5.20	72.45±4.40	76.10±3.80	78.90±3.20
Swin3D Liu et al. (2021)	61.80±6.10	68.30±5.00	74.20±4.10	77.60±3.50
ResNet He et al. (2016)	65.40±8.15	69.75±7.40	72.80±6.85	74.61±6.29
DenseNet Huang et al. (2017)	64.10±8.50	68.45±7.80	71.50±7.10	73.61±6.40

No architecture achieves strong performance in low-data regimes. At 30% training data (approximately 60 cases), the best model (ConvNeXt3D, 67.15%) barely exceeds random chance for a four-class task, and all architectures require the full dataset to approach their respective performance ceilings. From 70% to 100% of the training data, the ACC gains for the four backbones are +2.80% (ConvNeXt3D), +3.40% (Swin3D), +1.81% (ResNet), and +2.11% (DenseNet), with an average improvement of +2.53%. This demonstrates that existing architectures lack effective mechanisms to extract discriminative features from limited samples, highlighting data-efficient learning as a critical open challenge on this benchmark.

Transformer-based architectures exhibit severe data efficiency deficiencies. Swin3D exhibits an ACC gain of +15.80% from 30% to 100% of the training data, the largest among all models. In comparison, ConvNeXt3D, which benefits from stronger local inductive biases through pretrained convolutional features, achieves a gain of +11.75%, while ResNet and DenseNet show improvements of +9.21% and +9.51%, respectively. Notably, with only 60 training samples, Swin3D (61.80%) substantially underperforms ResNet (65.40%) and DenseNet (64.10%), but surpasses both at full data scale with 77.60%. This crossover phenomenon reveals that self-attention mechanisms, while powerful in representation capacity, lack the inductive biases necessary to learn from limited medical imaging data. This points to an important research direction: designing architectures or training strategies that combine the representational power of Transformers with the sample efficiency of convolutional priors for data-constrained clinical tasks.

Systematic increase in variance under low-data regimes. Under the 30% training data setting, all models exhibit substantially higher standard deviations compared to the 100% setting (e.g., ResNet: ±8.15 vs. ±6.29; DenseNet: ±8.50 vs. ±6.40), indicating reduced reliability of individual training runs at the 60-sample scale. This instability further confirms that existing architectures lack robust learning mechanisms for small-sample medical imaging scenarios, where training outcomes become highly sensitive to fold composition. Developing methods that yield stable performance under limited data remains an open challenge.

Summary. Overall, the results in Tab. 6 reveal systematic data efficiency deficiencies across all tested architectures, establishing DeepSite-MRI as a challenging testbed for research in data-constrained learning. The findings point to multiple open directions: designing architectures with stronger inductive biases for volumetric medical data, developing training strategies that improve sample efficiency, and exploring semi-supervised or self-supervised approaches that can leverage unlabeled clinical scans. Meanwhile, multi-center expansion of the dataset remains a complementary avenue for advancing this benchmark.

Table 7: Distribution of DeepSite-MRI cases across MRI systems and pathological categories.

MRI System	Cases	Glioma	PCNSL	Germinoma	Rare
Philips Ingenia CX	232	138	68	16	10
Siemens Prisma	11	5	4	2	0
GE SIGNA Premier	3	3	0	0	0
Philips Ingenia	3	3	0	0	0
Total	**249**	**149**	**72**	**18**	**10**

G Scanner Distribution

DeepSite-MRI is a multi-center cohort acquired using multiple MRI systems from three manufacturers. Among the 249 patients with complete scanner information, four MRI systems were represented. The detailed scanner-by-pathology distribution is shown in Table 7.

Thus, the current benchmark incorporates acquisition variability across multiple centers, manufacturers, and scanner systems. However, the strong concentration of cases on the Philips Ingenia CX means that the observed scanner heterogeneity should not be interpreted as fully representative of cross-center acquisition diversity. Differences in signal intensity, spatial resolution, contrast-agent timing, reconstruction algorithms, and scanning protocols across centers may still introduce domain shifts. DeepSite-MRI therefore provides a multi-center, multi-scanner benchmark while leaving room for further expansion to broader populations, institutions, and acquisition protocols.

H Dataset Value and Open Research Directions

Transferability of the Annotation Methodology. The annotation paradigm embodied by the TSP pipeline (zero-shot VLM coarse localization + MLP refinement + expert review) offers a reusable methodology for data-scarce medical imaging domains. In scenarios where conventional semi-automatic tools cannot be initialized due to the lack of in-domain pretrained models, TSP demonstrates an alternative path: leveraging the zero-shot localization capability of general-purpose VLMs to generate initial spatial priors without any domain adaptation, then training a lightweight MLP on a small number of samples to learn domain-specific refinement mappings. This two-stage design decomposes the annotation problem into "coarse localization" and "boundary refinement," with each subproblem addressed by the most suitable tool. Crucially, the coarse localization stage requires no training data (zero-shot), while the refinement stage needs only a small set of annotated samples, making the entire pipeline viable under extreme data scarcity. This paradigm is potentially transferable to dataset construction for other rare pathologies (e.g., rare dermatological or fundus lesions) or anatomical regions with prohibitively high annotation costs (e.g., spinal cord, skull base), providing the community with a low-cost, scalable annotation infrastructure.

Long-Tailed Recognition and Class-Imbalanced Learning. The natural class distribution of DeepSite-MRI, with Glioma comprising 59.8% of cases and the Rare category representing only 4.0% (10 instances), reflects real-world clinical incidence and constitutes an authentic long-tailed recognition benchmark. Unlike artificially subsampled imbalanced datasets, the scarcity of Rare cases here is medically irreducible: these tumors are genuinely rare, and no data collection effort can overcome this epidemiological reality. DeepSite-MRI therefore provides a clinically grounded testbed for long-tailed learning strategies, including class-reweighting, few-shot learning, meta-learning, and synthetic data augmentation via generative models, all in a setting where the cost of misclassifying a rare subtype has direct clinical consequences.

Data-Efficient and Transfer Learning. The constrained scale of DeepSite-MRI (249 cases) is not merely a limitation to be overcome; rather, it reflects the practical ceiling of data availability for rare, biopsy-gated medical imaging tasks. As such, DeepSite-MRI serves as a principled benchmark for data-efficient learning research, including transfer learning from large-scale pretrained models, few-shot adaptation, domain generalization from hemispheric tumor datasets, and active learning strategies for maximizing annotation utility. Our scaling analysis (Fig. 4) provides a baseline

characterization of each architecture's data efficiency curve and establishes a quantitative foundation for direct comparison by future data-efficient methods.

Radiomic Feature Discovery and Biomarker Research. From a translational medicine perspective, DeepSite-MRI offers the first public resource for radiomic biomarker discovery in deep-seated intracranial tumors. The combination of histopathologically confirmed subtype labels, full-cohort demographic metadata (age, sex), and high-fidelity 3D T1-CE volumes enables researchers to investigate: (i) which MRI-derived features are most discriminative between subtypes that appear visually similar (e.g., PCNSL vs. Glioma in the thalamus); (ii) whether radiomic signatures correlate with patient age or tumor location within the deep-seated region; and (iii) whether features extracted from coarse versus refined spatial priors carry differential prognostic information. Such analyses directly serve neuroradiologists and neurosurgeons seeking imaging biomarkers to reduce biopsy dependence in clinical practice.

Clinical Decision Support and AI-Assisted Biopsy Planning. For clinicians and clinical AI developers, DeepSite-MRI directly addresses the biopsy planning workflow. By enabling preoperative subtype prediction from T1-CE MRI, models trained on this benchmark could inform target selection for stereotactic biopsy and reduce the risk of inconclusive sampling caused by suboptimal target localization. More broadly, the benchmark supports research on AI-assisted differential diagnosis of deep-seated lesions, a clinically critical task where radiological differentiation among Glioma, PCNSL, and Germinoma is known to be unreliable without tissue confirmation. DeepSite-MRI is the first public resource that grounds this clinical problem in histopathologically verified labels, enabling rigorous validation of clinical decision support systems before prospective deployment.

In summary, DeepSite-MRI is designed not as a single-task dataset but as a shared infrastructure resource: its biopsy-confirmed labels, multi-level spatial annotations, full demographic metadata, and clinically grounded class distribution collectively support a spectrum of research directions that span computational methodology and clinical translation. We invite the community to build upon this foundation.