Evaluating SAM3 and MedGemma for Brain Tumor MRI: Zero-Shot Segmentation, Linear Probing, and Multimodal Diagnosis

Tamara Kostova, Ivan Kitanovski, Monika Simjanoska Misheva, and Kostadin Mishev
Faculty of Computer Science and Engineering
FINKI, Ss. Cyril and Methodius University
Skopje, North Macedonia
tamara.kostova.1@students.finki.ukim.mk {ivan.kitanovski, monika.simjanoska, kostadin.mishev}@finki.ukim.mk

***Abstract*—This paper presents a series of experiments applying two recent foundation models—SAM3 [1], a text-prompted image/video segmentation model, and MedGemma 1.5 4B [2], a multimodal medical vision–language model—to brain tumor analysis tasks. We first evaluate SAM3 in a zero-shot setting on all 125 patients from the BraTS 2021 dataset [3], observing limited performance (Dice $= 0.189$, IoU $= 0.124$, Sensitivity $= 0.397$). To address this, we apply *linear probing*: a single linear layer is trained on frozen SAM3 image-encoder features, substantially improving segmentation quality to Dice $= 0.836$ (pixel-level aggregate) and 0.801 (per-case mean, 95% CI $[0.770, 0.832]$), a gain of $+0.647$ over the zero-shot baseline without modifying the backbone. Independently, we evaluate MedGemma 1.5 4B on a multi-dataset brain MRI benchmark [4]–[7] covering five classification tasks: imaging modality, MRI sequence, image plane, binary diagnosis, and tumor subtype. Finally, we introduce a combined pipeline in which SAM3 provides spatial attention by overlaying a detected bounding box onto the full image before passing it to MedGemma. The pipeline improves tumor detection accuracy (85.1%→96.3%) and image-plane prediction (28.2%→51.6%), but introduces a critical normal-scan specificity collapse caused by SAM3 false-positive detections (67.1%→41.3%).**

***Index Terms*—brain tumor segmentation, foundation models, SAM3, MedGemma, BraTS, linear probing, vision–language models, MRI classification**

I. Introduction

Brain tumor analysis from magnetic resonance imaging (MRI) is a clinically important yet challenging task, requiring both accurate spatial localization of abnormal tissue and reliable semantic interpretation of imaging findings. Recent foundation models have shown strong capabilities in zero-shot segmentation [1], [8] and multimodal medical reasoning [2]. However, their applicability to specialized neuroimaging without domain-specific adaptation remains unclear. Brain tumor MRI presents a particularly demanding test bed for transfer learning for reasons that are specific to this domain. Tumor boundaries are often subtle and infiltrative: high-grade gliomas blend imperceptibly with surrounding oedematous tissue, producing no sharp edge that a natural-image segmentation model can latch onto. Tumor appearance is highly heterogeneous—it varies across patients, tumor grades, and MRI sequences (T1, T1-contrast-enhanced, T2, FLAIR)—meaning the same underlying pathology produces markedly different signal intensities depending on the imaging context. Acquisition protocols further vary across institutions, introducing scanner-specific noise, field-strength differences, and slice-thickness variation that are absent from the natural-image training distributions of general-purpose foundation models. Methods that appear robust on natural images or broad medical benchmarks may therefore fail when applied directly to neuro-oncologic imaging.

Prior work has shown that general-purpose segmentation models often underperform in the medical domain. SAM [8] and its derivatives struggle with the low-contrast boundaries and noise characteristics of clinical imaging, motivating targeted adaptations such as SAM-Adapter [9], MedSAM [10], SAMed [11], SAM-Med3D [12], and RefSAM3D [13]. In contrast to these approaches, we do *not* fine-tune the segmentation backbone. Instead, we treat SAM3 as a frozen feature extractor and train only a minimal linear head on top of its image-encoder representations. This linear probing setting [14] provides a direct test of whether the pre-trained representation already contains task-relevant information for brain tumor segmentation. This distinction is important because weak zero-shot predictions do not necessarily imply weak underlying representations. A model may fail at direct mask generation while still encoding features that are highly informative for downstream segmentation. Linear probing therefore provides a useful diagnostic tool for disentangling output-level failure from representation-level transferability.

Beyond segmentation, brain MRI analysis also requires semantic understanding, including recognition of imaging modality, MRI sequence, anatomical plane, and tumor-related diagnostic categories. Vision–language models (VLMs) such as MedGemma [2] offer a promising framework for such tasks, but their performance on brain MRI has not yet been systematically characterized across diverse classification settings. Evaluating VLMs on brain MRI is also nontrivial, as the task requires sensitivity to fine-grained visual properties such as imaging sequence, acquisition plane, and subtle pathological cues, rather than only broad semantic categorization. Moreover, it remains unclear whether spatial cues derived from segmentation models can improve downstream VLM-based reasoning. This

motivates a simple but important question: can spatial cues from a segmentation model improve downstream vision-language classification by directing attention to clinically relevant regions, or do localization errors instead amplify diagnostic mistakes?

In this paper, we investigate the transferability of two recent foundation models, SAM3 and MedGemma 1.5 4B, to brain tumor MRI analysis without full model fine-tuning. We first evaluate SAM3 in a zero-shot segmentation setting on the BraTS benchmark and show that, although direct zero-shot performance is inadequate, the model exhibits partial sensitivity to tumor regions. We then demonstrate that frozen SAM3 features support strong segmentation performance when paired with a simple linear head, indicating that useful tumor-relevant representations are present even when zero-shot predictions fail. Independently, we benchmark MedGemma on a multi-task brain MRI classification setting spanning imaging modality, MRI sequence, anatomical plane, tumor detection, and tumor subtype recognition. Finally, we study a combined SAM3–MedGemma pipeline in which segmentation-derived spatial cues are used to guide VLM inference, highlighting both its benefits and its failure modes. Taken together, our experiments show that transferability of foundation models in brain MRI must be assessed at multiple levels: direct zero-shot behavior, representation quality under lightweight adaptation, and robustness when models are combined into multi-stage pipelines.

Our contributions are fourfold:

1) We evaluate SAM3 on the BraTS brain tumor segmentation benchmark in a zero-shot setting and show that its direct performance is insufficient for clinical use, despite non-trivial sensitivity to tumor regions.
2) We show that frozen SAM3 representations can be effectively exploited through linear probing, achieving Dice $= 0.836$ at the pixel level and a per-case mean Dice of 0.801 using only a single 1×1 convolutional head, without updating the backbone.
3) We present a systematic benchmark of MedGemma 1.5 4B on brain MRI classification tasks, including imaging modality, MRI sequence, anatomical plane, tumor detection, and tumor subtype recognition.
4) We propose and analyze a segmentation-guided SAM3–MedGemma pipeline, characterizing the conditions under which spatial cues improve downstream predictions and the failure modes that arise when false-positive detections are propagated.

II. Related Work

A. SAM and Its Adaptations for Medical Imaging

The Segment Anything Model (SAM) [8] demonstrated strong zero-shot performance on natural images but degrades on medical data due to low contrast, ambiguous boundaries, and domain shift. Numerous adaptations have been proposed: SAM-Adapter [9] injects domain-specific features into SAM's prompt encoder; MedSAM [10] fine-tunes the full model on a large curated collection of medical image–mask pairs; SAMed [11] applies low-rank adaptation for abdominal CT; SAM-Med3D [12] extends the architecture to volumetric inputs; and RefSAM3D [13] introduces language-guided prompting for 3-D data. SAM3 [1] adds open-vocabulary text prompting without volumetric fine-tuning.

In contrast to prior medical SAM adaptations, our goal is not to improve segmentation through backbone fine-tuning or architectural redesign. Instead, we study whether a recent promptable segmentation foundation model can transfer to brain tumor MRI in two lightweight settings: direct zero-shot inference and linear probing of frozen encoder features. This positions our work as an evaluation of representational transferability rather than a new adaptation method.

B. Transfer Learning and Linear Probing for Medical Segmentation

Transfer learning is widely used in medical image analysis, where limited labeled data often makes training large models from scratch impractical. In this context, linear probing, which exploits training only a linear prediction layer on top of frozen backbone features, has become a standard tool for measuring representation quality [14]. Large-scale pre-training yields image encoders that provide strong initialisations for medical segmentation [10], and linear probing has shown that such features retain task-relevant structure even without fine-tuning [14]. Few-shot or universal segmentation methods, such as UniverSeg [15], instead generalise across tasks by conditioning on support image–label pairs via cross-attention. In contrast, our work investigates whether frozen foundation-model features alone suffice for segmentation via linear probing, without any task-specific support examples.

C. Medical Vision-Language Models

Vision-language models (VLMs) have recently emerged as a promising framework for medical image understanding by combining visual perception with language-conditioned reasoning. Early and recent medical vision-language models (VLMs) include LLaVA-Med [16], which adapts LLaVA [17] to biomedical image-text pairs; BioViL-T [18], which models temporal alignment across longitudinal chest X-ray and radiology report sequences; and Med-Flamingo [19], a few-shot adaptation of Flamingo [20]. These models suggest that multimodal pre-training can support flexible medical question answering, report-style reasoning, and structured prediction.

However, most existing evaluations of medical VLMs focus on modalities and tasks such as chest radiography, pathology, ophthalmology, or general biomedical visual question answering, while brain MRI remains relatively underexplored. Neuroimaging poses additional challenges because successful interpretation depends on fine-grained distinctions among MRI sequences, anatomical planes, and subtle pathological patterns. MedGemma 1.5 4B [2] extends the Gemma 2 family with medical pre-training and instruction tuning for multimodal clinical tasks. In this work, we evaluate MedGemma on a diverse set of brain MRI classification tasks and further study

whether segmentation-derived spatial cues can improve its downstream predictions.

D. Segmentation-Guided Multimodal Reasoning

Combining spatial localization with semantic reasoning is an increasingly important direction in medical AI, motivating pipelines in which a segmentation or detection model first identifies candidate regions of interest and a second model then performs classification, report generation, or vision-language reasoning on the guided input [21]–[25].

Region-guided multimodal methods indicate that explicit spatial grounding can improve clinically relevant prediction by reducing background clutter and promoting tighter alignment between localized visual evidence and downstream textual or diagnostic outputs [21]–[23]. At the same time, these multi-stage pipelines remain vulnerable to error propagation: false-positive detections or imprecise localization at the first stage can steer the downstream classifier or vision–language model toward incorrect conclusions [24], [25]. Our SAM3–MedGemma pipeline is an instance of this segmentation-guided design. Rather than proposing a new multimodal architecture, we evaluate a simple interface in which segmentation-derived spatial cues guide downstream VLM inference, and we characterize both its benefits and its failure modes for brain MRI classification.

E. Brain Tumor Segmentation Benchmarks

The BraTS challenge series [3], [26] provides a standardized platform for tumor segmentation evaluation. Fully supervised state-of-the-art methods achieve Dice scores of 0.85–0.92 on the enhancing tumor sub-region. Domain-adapted SAM variants demonstrate that fine-tuning substantially narrows the gap with specialist models. A recent benchmark by Ludwig et al. [27] directly compared SAM, SAM 2, MedSAM [10], SAM-Med3D [12], and nnU-Net on the BraTS 2023 adult glioma dataset across multiple prompt qualities. MedSAM achieves Dice $= 0.685$ with high-quality bounding-box prompts, and SAM-Med3D achieves Dice $= 0.815$ with 10-point prompts; however, both models benefit from data leakage, as BraTS data was included in their respective training sets—a confound explicitly noted by the authors. When evaluated on the BraTS 2023 pediatric subset (no leakage), their scores drop to 0.572 and 0.814 respectively. Both require substantial domain-specific supervision and task-aligned training data. In contrast, our linear probe achieves Dice $= 0.836$ on BraTS 2021 using only frozen SAM3 backbone features and a single 1×1 convolutional head—without backbone modification and without any BraTS data in the upstream training—representing a minimal-supervision alternative that matches or exceeds these domain-adapted models through representation reuse alone.

III. Background

A. SAM3

SAM3 [1] extends the Segment Anything Model (SAM) [8] family with improved promptability and open-vocabulary segmentation capabilities. Like SAM, it follows an encoder–decoder design consisting of a high-capacity image encoder, a prompt encoder, and a lightweight mask decoder. The image encoder produces a dense spatial feature map, while the prompt encoder conditions the model on input prompts (e.g., text, points, or boxes), and the mask decoder generates segmentation masks via cross-attention between prompt embeddings and image features.

Compared to the original SAM, SAM3 introduces two key differences relevant to this work. First, it incorporates *text-conditioned prompting*, enabling segmentation from free-form natural language descriptions (e.g., "enhancing tumor region") rather than requiring explicit spatial prompts. Given a text query, SAM3 produces pixel-level binary masks with associated confidence scores, making it an attractive zero-shot candidate for medical imaging tasks where domain-specific training data is scarce. Second, SAM3 is trained on a broader and more diverse set of image–text pairs, improving its open-vocabulary generalisation beyond object-centric natural images. Unlike SAM-Med3D or other medical adaptations, however, SAM3 remains a 2-D model and is not explicitly trained on volumetric medical data.

B. MedGemma

MedGemma 1.5 (`google/medgemma-1.5-4b-it`) [2] is a 4-billion parameter instruction-tuned vision-language model built on the Gemma 3 architecture. Pre-trained on radiology, pathology, and ophthalmology image–text pairs, it accepts interleaved image–text inputs and produces structured textual outputs, making it well-suited for multi-attribute MRI classification via structured JSON prompting.

C. Datasets

BraTS 2021. The BraTS 2021 benchmark [3], [26], [28] consists of multi-parametric brain MRI volumes, including FLAIR, T1, T1ce, and T2 sequences, with expert voxel-level tumor segmentation annotations. We use a cohort of 125 patients for two purposes: (i) zero-shot evaluation of SAM3 and (ii) the held-out test set for the linear-probe experiments (Section V-A). The linear probe is trained on the BraTS 2020 training and validation splits. To avoid patient overlap between training and testing, we verified across official dataset mappings that no PatientID or TCGA identifier is shared between the BraTS 2020 training data and the 125-patient BraTS 2021 test cohort [29]–[31]. The test cohort originates from UCSF-PDGM, CPTAC-GBM, and IvyGAP collections that post-date the construction of BraTS 2020.

MedGemma benchmark. For classification, we assemble a benchmark from four publicly available (Table I) brain MRI datasets: the Figshare brain tumor dataset [4], containing glioma, meningioma, and pituitary tumor classes; Br35H [5], containing binary tumor-versus-normal labels; and two curated public datasets with 17 [7] and 44 [6] tumor subclasses, respectively. The 17-class and 44-class datasets also contain filename-encoded metadata, including MRI sequence (e.g., `T1`, `T2`) and anatomical plane (e.g., `axial`), which we parse to

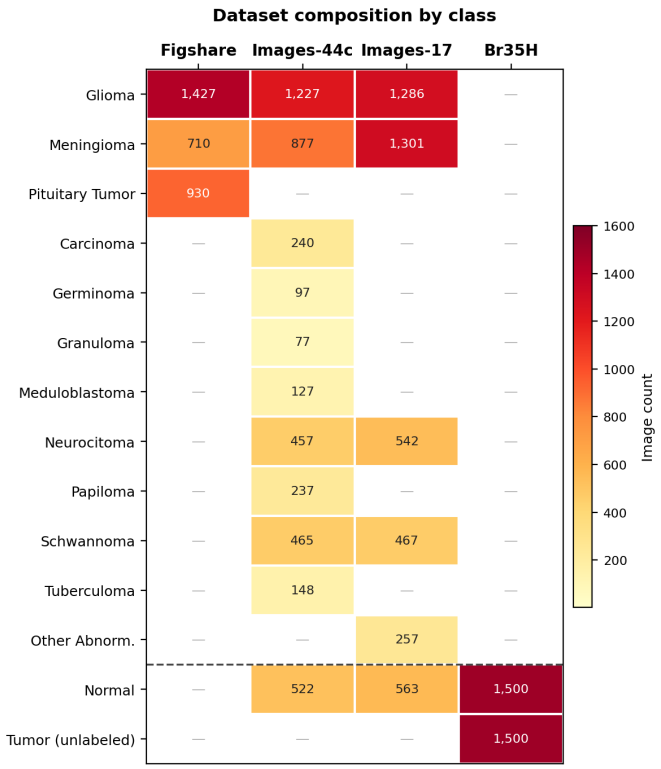


Fig. 1. Class distribution across the four benchmark datasets. Cell values show image counts; dashes indicate classes absent from a given dataset. The dashed line separates tumor subtype classes (above) from normal and coarse-label tumor images (below). Detailed counts per sequence and plane are given in Table I.

construct ground-truth labels for the sequence-classification and plane-classification tasks. Because all included images are brain MRI scans, the modality label is uniformly assigned as MRI across all four datasets.

Fig. 1 visualises the per-class distribution across all four datasets. Sequence classification is evaluated on all datasets except Br35H (n=10,393), which carries no sequence annotation. Plane classification is evaluated on Images-17 and Br35H only (n=5,852), as Figshare and Images-44c have no plane metadata.

D. Metrics

We use different evaluation metrics depending on whether the task is pixel-level segmentation or image-level classification.

1) Segmentation Metrics: For SAM3 zero-shot segmentation and linear probing (Sections IV-A and VI-A), performance is measured with the following pixel-level metrics. Let P denote the set of predicted foreground pixels and G the set of ground-truth foreground pixels.

Dice Score (F1). The harmonic mean of precision and recall over pixel sets:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}. \tag{1}$$

Dice ranges from 0 (no overlap) to 1 (perfect overlap) and is the primary metric for segmentation quality, as it penalises both false positives and false negatives equally.

Intersection over Union (IoU / Jaccard Index).

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|}. \tag{2}$$

IoU is a stricter overlap measure than Dice; the two are related by $\text{IoU} = \text{Dice}/(2 - \text{Dice})$.

Sensitivity (Recall). The fraction of true tumor pixels that are correctly detected:

$$\text{Sensitivity} = \frac{|P \cap G|}{|G|}. \tag{3}$$

Sensitivity is computed at the pixel level and aggregated across slices. High sensitivity with low Dice indicates that the model covers the tumor but generates excessively large masks, a pattern observed in SAM3 zero-shot predictions.

Precision. The fraction of predicted foreground pixels that truly belong to the tumor:

$$\text{Precision} = \frac{|P \cap G|}{|P|}. \tag{4}$$

Precision is reported for both zero-shot and linear probe, where it complements sensitivity to characterise mask compactness.

Pixel-level accuracy. The fraction of all pixels (foreground and background) correctly classified. Although easy to interpret, this metric is dominated by the large background class in MRI and is therefore reported only as a secondary indicator alongside Dice and IoU.

Aggregation. All segmentation metrics are computed at two levels of granularity.

Pixel-level aggregates pool all foreground and background pixels across every slice in the test set before computing a single global metric; this weights each pixel equally and is primarily reported for the zero-shot and linear-probe comparisons.

Per-case means compute the metric independently for each patient (averaging over that patient's slices) and then average across patients; this treats every case equally regardless of volume size and is reported alongside 95% confidence intervals computed using a t-distribution. Where the two levels diverge—as seen in the linear probe results (pixel-level Dice $= 0.836$ vs. per-case mean $= 0.801$)—the difference reflects patient-level heterogeneity rather than methodological inconsistency.

2) Classification Metrics: **Accuracy.** The fraction of images assigned the correct label, reported both overall and per class:

$$\text{Accuracy} = \frac{\text{correct predictions}}{\text{total predictions}}. \tag{5}$$

Per-class accuracy is reported separately to expose class-level biases such as the glioma dominance observed in subtype prediction.

Wilson Score Confidence Interval. Because accuracy is a proportion, we compute 95% confidence intervals using the Wilson score method [32], which remains well-calibrated for both small samples and proportions near 0 or 1, unlike the normal approximation interval.

TABLE I

BENCHMARK DATASET COUNTS AND DISTRIBUTION ACROSS MODALITY, MRI SEQUENCES, AND ACQUISITION PLANE. ALL IMAGES ARE MRI SCANS. †OF THE 14,957 IMAGES, 13,390 WERE MATCHED TO SAM3 PIPELINE OUTPUTS AND USED FOR PAIRED COMPARISONS (TABLES VII AND VI).

Identity			Count	Modality		Sequence (MRI)				Plane	
Dataset	**Class**	**Subclass**		**CT**	**MRI**	**FLAIR**	**T1C+**	**T1**	**T2**	**Axial**	**Sagittal**
Figshare	Tumor	Glioma	1,427	0	1,427	0	1,427	0	0	0	0
		Meningioma	710	0	710	0	710	0	0	0	0
		Pituitary Tumor	930	0	930	0	930	0	0	0	0
	Dataset Total		*3,067*	*0*	*3,067*	*0*	*3,067*	*0*	*0*	*0*	*0*
Images-44c	Tumor	Carcinoma	240	0	240	0	112	66	73	0	0
		Germinoma	97	0	97	0	40	27	33	0	0
		Glioma	1,227	0	1,227	0	464	382	372	0	0
		Granuloma	77	0	77	0	31	30	17	0	0
		Meduloblastoma	127	0	127	0	67	23	41	0	0
		Meningioma	877	0	877	0	369	272	233	0	0
		Neurocitoma	457	0	457	0	223	130	104	0	0
		Papiloma	237	0	237	0	108	66	63	0	0
		Schwannoma	465	0	465	0	194	148	123	0	0
		Tuberculoma	148	0	148	0	84	28	33	0	0
	Normal	—	522	0	522	0	0	251	271	0	0
	Dataset Total		*4,474*	*0*	*4,474*	*0*	*1,692*	*1,423*	*1,363*	*0*	*0*
Images-17	Tumor	Glioma	1,286	0	1,286	0	508	430	346	1,286	0
		Meningioma	1,301	0	1,301	0	625	345	329	1,301	0
		Neurocitoma	542	0	542	0	261	169	112	542	0
		Schwannoma	467	0	467	0	194	153	123	467	0
	Other Abnormalities	—	257	0	257	0	48	152	57	257	0
	Normal	—	563	0	563	0	0	272	291	563	0
	Dataset Total		*4,416*	*0*	*4,416*	*0*	*1,636*	*1,521*	*1,258*	*4,416*	*0*
Br35H	Tumor	—	1,500	0	1,500	0	0	0	0	1,500	0
	Normal	—	1,500	0	1,500	0	0	0	0	1,500	0
	Dataset Total		*3,000*	*0*	*3,000*	*0*	*0*	*0*	*0*	*3,000*	*0*
Grand Total†			**14,957**	**0**	**14,957**	**0**	**6,395**	**2,944**	**2,621**	**7,416**	**0**

McNemar's Test. To assess whether the combined SAM3 + MedGemma pipeline significantly changes classification accuracy relative to standalone MedGemma, we use McNemar's test [33] on paired correct/incorrect outcomes. McNemar's test is appropriate here because the two systems are evaluated on the same set of images, making outcomes paired; it focuses specifically on the discordant cases (images where one system is correct and the other is wrong) rather than the concordant ones, providing a sensitive test of directional improvement or degradation. All reported p-values are two-sided; a threshold of $\alpha = 0.05$ is used for significance.

IV. BASELINE RESULTS

This section reports the zero-shot performance of SAM3 and MedGemma 1.5 4B applied directly to brain MRI without any fine-tuning, adaptation, or inter-model combination.

A. SAM3 Zero-Shot Segmentation on BraTS 2021

Setup. We evaluated SAM3 zero-shot on T1ce volumes from 125 BraTS 2021 patients. Axial slices in the range 50–130 were extracted, normalised to $[0, 1]$, and converted to RGB. Multiple text prompts were tested; *"enhancing tumor region"* consistently yielded the best scores. Predicted masks were compared against ground-truth segmentations using Dice, Intersection-over-Union (IoU), and sensitivity.

Results. SAM3 produced non-empty masks on all 125 cases. Results are shown in Table II. Values are pixel-level aggregates with 95% Confidence Interval (CI) computed over the full slice population.

TABLE II

SAM3 ZERO-SHOT SEGMENTATION ON BRATS 2021 ($n = 125$ CASES). PIXEL-LEVEL AGGREGATE WITH 95% CIS

Metric	Value	95% CI
Dice Score	0.189	[0.184, 0.193]
IoU	0.124	[0.120, 0.128]
Sensitivity	0.397	[0.389, 0.405]

SAM3 shows some capacity to localise enhancing tumor tissue, as reflected by a sensitivity of 0.397, indicating that the model captures a non-trivial portion of tumor pixels, but still misses the majority of the lesion. However, precision is poor, yielding low Dice and IoU. Pixel-level precision is extremely low (0.097), strongly indicating systematic over-segmentation, as evidenced by the generation of overly large masks that include many false positives. The relatively narrow confidence intervals reflect the large number of aggregated pixels rather than low variability across patients.

At the per-case level, sensitivity falls to 0.319 (95% CI $[0.293, 0.345]$) and precision to 0.132 (95% CI $[0.111, 0.152]$), as per-case averaging weights every patient equally regardless of tumor volume, whereas the pixel-level aggregate is dominated by large tumors with high absolute true-positive

counts.

Performance is highly variable across cases: the per-case mean Dice is 0.186 (std 0.096, median 0.163), with near-zero scores on small or diffuse tumors pulling the distribution leftward. Qualitative inspection suggests that performance may be better on well-defined lesions, while small or diffuse tumors yield poor predictions.

These results indicate that SAM3 zero-shot performance is insufficient for reliable standalone clinical use in brain MRI segmentation under zero-shot settings.

B. Standalone MedGemma Evaluation

Setup. MedGemma 1.5 4B was queried via a structured neuroradiology system prompt at temperature 0 through a vLLM endpoint. We evaluated five classification tasks: imaging modality, MRI sequence (T1, T1c+, T2), image plane (axial), binary diagnosis (normal / tumor), and tumor subtype (11 classes). All results are reported in Table III; accuracy values are accompanied by Wilson score 95% CIs.

TABLE III
STANDALONE MEDGEMMA 1.5 4B CLASSIFICATION ACCURACY (ZERO-SHOT)

Task	Subclass	Correct / Total	Acc.
Modality	MRI (all)	12,114 / 13,390	90.5%
Sequence	T1	340 / 2,439	13.9%
	T1c+	1,548 / 5,706	27.1%
	T2	662 / 2,248	29.4%
	Overall	2,550 / 10,393	*24.5%*
Plane	Axial	1,653 / 5,852	28.2%
Diagnosis	Normal	1,735 / 2,584	67.1%
	Tumor	9,192 / 10,806	85.1%
	Overall	10,927 / 13,390	*81.6%*
Subtype	Glioma	2,752 / 3,940	69.8%
	Meningioma	84 / 1,586	5.3%
	Schwannoma	25 / 928	2.7%
	Pituitary	0 / 930	0.0%
	Overall	2,862 / 9,309	*30.7%*

MedGemma is strong on high-level tasks (modality: 90.5%, binary diagnosis: 81.6%) but weak on fine-grained radiological attributes (sequence: 24.5%, plane: 28.2%, subtype: 30.7%). Weak sequence and plane performance is consistent with these tasks requiring low-level signal-intensity integration rather than semantic content recognition. A strong glioma bias dominates subtype prediction (69.8%), while all other subtypes score near zero, suggesting that prediction is driven by a frequency prior from training data rather than discriminative morphological reasoning. Average diagnosis confidence was high ($\mu = 0.948$); however, we note that high confidence does not imply calibration in the formal sense [34]—expected calibration error (ECE) measurements would be required to assess calibration, and we leave this to future work.

V. METHODOLOGY

Building on the baselines established in Section IV, we introduce three experimental protocols: linear probing of frozen SAM3 features, a generalisation study to other lesion types, and a segmentation-guided combined pipeline.

A. Linear Probing on Frozen SAM3 Features

Motivation. Linear probing trains a single linear layer on top of a frozen backbone, leaving all pre-trained weights unchanged. This serves as a principled evaluation of representational quality: if a linear classifier can extract task-relevant information from the fixed features, the backbone encodes sufficient spatial and textural statistics for the target task, even without domain-specific pre-training [14]. Crucially, weak zero-shot predictions do not necessarily imply weak underlying representations—a model may fail at direct mask generation while still encoding features that are highly informative for downstream segmentation.

Setup. We used the frozen SAM3 image encoder as a feature extractor and trained a lightweight `LinearProbe` segmentation head (a single 1×1 convolutional layer) on top. T1ce volumes were preprocessed to 1008×1008 px to match SAM3's input resolution. Features were pre-computed once and stored in HDF5 format. Training used the official BraTS 2020 training and validation sets (Adam, $\mathrm{lr} = 0.001$, 20 epochs) and was evaluated on the held-out 125-patient BraTS 2021 test set—the same cohort used in Section IV-A—enabling direct comparison.

Statistical analysis. To assess whether linear probing significantly improves per-case Dice over the zero-shot baseline, we performed a paired non-parametric test. Per-slice Dice values for zero-shot SAM3 were aggregated to per-case means and aligned with the linear-probe per-case Dice via the BraTS 2021 case ID, yielding $n = 125$ matched patients. Normality of the paired differences (probe$-$zero-shot) was evaluated with the Shapiro–Wilk test. Because normality was rejected (Section VI-A1), we used a one-sided Wilcoxon signed-rank test with alternative hypothesis H_1: $\mathrm{Dice}_{\mathrm{probe}} > \mathrm{Dice}_{\mathrm{zeroshot}}$. All statistics were computed in Python using NumPy, SciPy, and pandas.

B. Generalisation to Other Neuroimaging Tasks

To probe the limits of frozen SAM3 representations beyond brain tumor segmentation, we applied the same linear probing protocol to two additional tasks: MS lesion segmentation (MSLesSeg dataset) [35] and ischaemic stroke segmentation (ISLES 2022) [36].

Training setup. Both probes used the same 1×1 convolutional head as the BraTS experiment, optimised with AdamW ($\mathrm{lr} = 10^{-3}$, weight decay $= 10^{-4}$) over 20 epochs with cosine annealing ($\eta_{\min} = 0.01 \cdot \eta_0$). Unlike the BraTS probe, which uses plain cross-entropy, the MS and stroke probes use a combined loss $\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(\mathbf{w}) + 0.5\,\mathcal{L}_{\mathrm{Dice}}$, where $\mathbf{w}$ are inverse-frequency class weights derived from the dataset. The *empty-slice ratio* (fraction of lesion-free slices included during training) was set to 1.0 (all slices retained) to match the true class distribution; retaining all slices improved MS Dice from 0.149 to 0.263 and stroke Dice from 0.224 to 0.408 compared to an initial ratio of 0.3. All results reported in Section VI-B use

the corrected (1.0) ratio. Inspection of SAM3 feature activations confirmed higher mean responses in lesion regions, justifying the efficacy of a simple linear probe.

For **MS lesion segmentation**, we trained on 53 MSLesSeg [35] patients (training split) and evaluated on 22 held-out test cases (4,004 slices total, including 2,253 empty slices).

For **stroke lesion segmentation**, we trained on 75 ISLES 2022 patients (DWI modality) and evaluated on the remaining 25.

C. SAM3 + MedGemma Combined Pipeline

Pipeline design. We designed a two-stage pipeline illustrated in Fig. 2. SAM3 first generates a tumor-region proposal from the input image. When a non-trivial mask is detected (mask pixel fraction $> \theta$, where $\theta = 0$ for the base pipeline), its axis-aligned bounding box is drawn in red onto the *original full image*, and this annotated image is passed to MedGemma. The full image is retained so that MedGemma preserves global anatomical context while receiving an explicit spatial cue about the region of interest. When SAM3 produces no mask, the unmodified full image is used directly. Fig. 3 illustrates both a correct bounding box localisation and a representative false-positive detection.

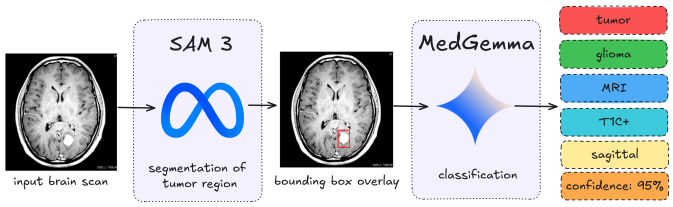
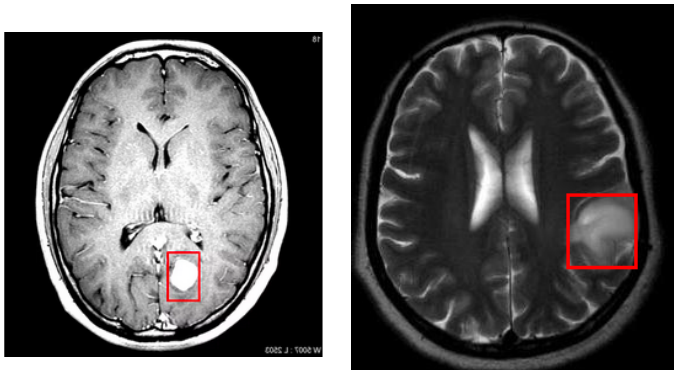


Fig. 2. SAM3 + MedGemma combined pipeline. SAM3 generates a tumor proposal mask from the input image. If a non-trivial mask is found (pixel fraction $> \theta$), the bounding box is overlaid on the *full* image before passing to MedGemma; otherwise, the unmodified image is used. This dual path ensures MedGemma retains global anatomical context regardless of SAM3's output.



(a) Correct bounding box localisation

(b) False-positive detection

Fig. 3. SAM3 bounding box behaviour. (a) Correct localisation provides a useful spatial prior. (b) False-positive detection on normal MRI can mislead downstream models, leading to incorrect classification.

VI. Results

A. Linear Probe Performance on BraTS 2021

Table IV compares zero-shot and linear-probe performance on the 125-patient test set (mean ± 95% CI).

TABLE IV
SAM3 Zero-Shot vs. Linear Probe on BraTS 2021 ($n = 125$). Pixel-level unless noted.

Method	Dice	IoU	Sens.	Prec.
Zero-shot (pixel)	0.189 [.184,.193]	0.124 [.120,.128]	0.397 [.389,.405]	0.097*
Zero-shot (case)	(0.186) [.170,.203]	(0.088) [.078,.097]	(0.319) [.293,.345]	(0.132) [.111,.152]
Probe (pixel)	**0.836**	**0.719**	**0.796**	**0.881**
Probe (case)	**(0.801)** [.770, .832]	**(0.693)** [.662, .725]	**(0.771)** [.737, .806]	**(0.864)** [.838, .889]
Δ (pixel)	+0.647	+0.595	+0.399	+0.784
Δ (case)	+0.615	+0.605	+0.452	+0.732

* Pixel-level precision is computed as a single global ratio ($\sum \mathrm{TP}/(\sum \mathrm{TP} + \sum \mathrm{FP})$); confidence intervals are not reported because the metric is not derived from per-slice estimates.

The linear probe achieves Dice $= 0.836$, IoU $= 0.719$, sensitivity $= 0.796$, and precision $= 0.881$ at pixel level—a substantial improvement over the zero-shot baseline on every metric, and competitive with lightly supervised domain-adapted SAM variants. Pixel-level accuracy across all test slices is 99.3%; however, this metric is dominated by the large background class and therefore Dice and IoU provide a more informative measure of tumor delineation quality. These results demonstrate that the frozen SAM3 encoder already captures spatial and textural statistics sufficient for a single linear layer to learn tumor-specific delineation without any backbone modification.

Notably, the per-case Dice mean (0.801) is lower than the pixel-level aggregate (0.836), reflecting variability across patients. The higher median Dice (0.865) compared to the mean indicates a left-skewed distribution, where most cases are segmented well but a subset of difficult tumors (e.g., small or diffuse lesions) reduce the average. The per-case IoU mean of 0.693 (95% CI $[0.662, 0.725]$) similarly reflects this patient-level variability.

1) Statistical significance of probe vs. zero-shot: For each of the 125 BraTS 2021 patients, we computed a per-case Dice score for zero-shot SAM3 by averaging slice-level Dice over all axial slices and matched these to the per-case linear-probe Dice via case identifier. The resulting distributions had mean Dice 0.186 (median 0.163, std 0.096) for zero-shot and 0.801 (median 0.865, std 0.175) for the linear probe.

The Shapiro–Wilk test on the paired differences (probe minus zero-shot) strongly rejected normality ($p < 10^{-4}$), justifying a non-parametric paired test. A one-sided Wilcoxon signed-rank test with alternative hypothesis $\mathrm{Dice}_{\mathrm{probe}} > \mathrm{Dice}_{\mathrm{zeroshot}}$ yielded $W = 7863.0$ and $p = 1.98 \times 10^{-22}$, indicating that the improvement from linear probing is highly significant at conventional levels.

B. Generalisation to MS and Stroke Segmentation

MS Lesion Segmentation (MSLesSeg). The zero-shot SAM3 baseline, using the prompt *"white matter lesion"*, yields pixel-level Dice $= 0.052$. The linear probe improves this to Dice $= 0.263$ ($+0.210$), with IoU $= 0.151$ and accuracy $= 0.983$ (Table V).

Stroke Lesion Segmentation (ISLES 2022). The zero-shot SAM3 baseline (prompt: *"ischemic infarct"*) achieves per-case Dice $= 0.107$ (95% CI $[0.061, 0.154]$), median Dice $= 0.085$, and global pixel-level Dice $= 0.342$ (IoU $= 0.207$). The linear probe achieves test Dice $= 0.408$, IoU $= 0.256$, and accuracy $= 0.987$. The gap between validation best-Dice (0.449) and test Dice (0.408) suggests mild overfitting given the small dataset size.

TABLE V
LINEAR PROBE GENERALISATION TO MS AND STROKE SEGMENTATION.

Task	Method	Dice	IoU	Acc.
MS (MSLesSeg)	Zero-shot SAM3	0.052	0.027	0.999
	Linear Probe	**0.263**	**0.151**	0.983
Stroke (ISLES)	Zero-shot SAM3†	0.107	0.207	—
	Linear Probe	**0.408**	**0.256**	0.987

†Zero-shot Dice is a per-case mean; stroke IoU is reported at global pixel level.

Both tasks yield substantially lower probe performance than BraTS (Dice $= 0.836$), reflecting three compounding factors. First, both datasets are small (MSLesSeg: $\sim$75 patients; ISLES: 100 patients), limiting the linear probe's ability to learn reliable thresholds. Second, class imbalance is severe: MS lesions cover approximately 1–2% of voxels per slice, compared with $\sim$10% for enhancing tumor in T1ce, creating a $\sim$50:1 foreground-to-background ratio that overwhelms a single linear layer under standard inverse-frequency weighting. Third, SAM3 downsamples the input to a coarse feature map, which may subsume small MS plaques (2–5 mm) within a single feature vector, destroying the spatial detail needed for accurate delineation. These results suggest that frozen SAM3 features are most effective when lesions are large, high-contrast, and well-represented in the training set.

C. SAM3 + MedGemma Pipeline

Table VI reports the base pipeline result alongside the standalone baseline; a full sweep over $\theta \in \{0.10, 0.30, 0.50\}$ is deferred to future work (Section VIII).

TABLE VI
EFFECT OF SAM3 DETECTION THRESHOLD ON DIAGNOSIS ACCURACY

Configuration	Tumor Acc. (%)	Normal Acc. (%)
Standalone MedGemma (no SAM3)	85.1	67.1
Pipeline ($\theta = 0.00$)	96.3	41.3

Table VII compares the base pipeline ($\theta = 0$) against standalone MedGemma across all five tasks. Statistical significance is assessed by McNemar's test on paired correct/incorrect outcomes; matched n values per task range from 4,162 (plane) to 13,390 (modality).

TABLE VII
SAM3 + MEDGEMMA PIPELINE VS. STANDALONE MEDGEMMA (Δ IN PP; p FROM MCNEMAR'S TEST, PAIRED). BOTH SYSTEMS EVALUATED ON THE SAME n=10,119–13,390 MATCHED IMAGES PER TASK.

Task	Standalone (%)	Pipeline (%)	Δ	p
Modality	90.5	98.1	$+7.6$	< 0.001
Sequence	24.5	29.5	$+5.0$	< 0.001
Image Plane	28.2	51.6	$+23.4$	< 0.001
Diagnosis (overall)	81.6	87.0	$+5.4$	< 0.001
Diagnosis (tumor)	85.1	96.3	$+11.2$	< 0.001
Diagnosis (normal)	67.1	41.3	-25.8	< 0.001
Subtype (overall)	30.7	17.0	-13.7	< 0.001

Results are also visualised in Fig. 4. The pipeline substantially improves modality detection ($+7.6$ pp), image-plane prediction ($+23.4$ pp), and tumor detection ($+11.2$ pp). The large plane improvement likely results from the bounding box removing uninformative background, positioning anatomy more canonically for MedGemma. Three failure modes emerge:

1) **Normal-scan specificity collapse.** SAM3 incorrectly flagged tumor-like regions in 577 of 2,046 normal images (25.5%), of which 513 (88.9%) were subsequently misclassified as tumor, dropping normal accuracy from 67.1% to 41.3%. The flagged regions typically correspond to normal anatomical structures with locally elevated signal intensity, including choroid plexus, cerebral vessels, and sulcal margins.
2) **Subtype accuracy degradation.** Overall subtype accuracy fell from 30.7% to 17.0%. A key driver is that the bounding box caused MedGemma to output the generic label "tumor" (rather than a resolved subtype) on 4,260 pipeline-routed images (42.7% of tumor cases), collapsing fine-grained discrimination into coarse detection.
3) **Missed-tumor mislocation.** SAM3 failed to generate any mask on 1,018 of 9,978 tumor images (10.2%), passing unmodified images to MedGemma. Of these, 128 (12.6%) were then misclassified. The most common missed-tumor subtypes were meningioma (predicted normal in the majority of failures) and germinoma, consistent with low-contrast enhancement on T1ce.

1) Per-image influence analysis: To quantify SAM3's directional influence at the individual image level, we cross-referenced standalone and pipeline predictions for all matched images ($n = 10{,}119$, diagnosis task). Each image falls into one of four outcomes: both systems correct, both wrong, SAM3 helped (standalone wrong $\rightarrow$ pipeline correct), or SAM3 hurt (standalone correct $\rightarrow$ pipeline wrong). Results are summarised in Fig. 5.

The majority of images were handled correctly by both systems (7,510; 74.2%). SAM3 helped on 1,133 images (11.2%) and hurt on 709 (7.0%), yielding a **net benefit of $+424$ images** in favour of the pipeline. Both systems failed on 767 images

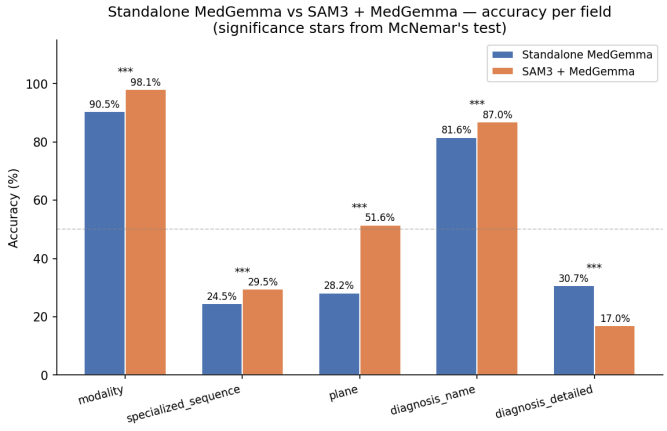


Fig. 4. Accuracy comparison across five classification tasks for standalone MedGemma (blue) vs. the SAM3 + MedGemma pipeline (orange). Error bars indicate Wilson-score 95% CIs. The pipeline improves modality, sequence, plane, and tumor-detection accuracy, but degrades normal-scan specificity and subtype accuracy due to false-positive region propagation from SAM3.

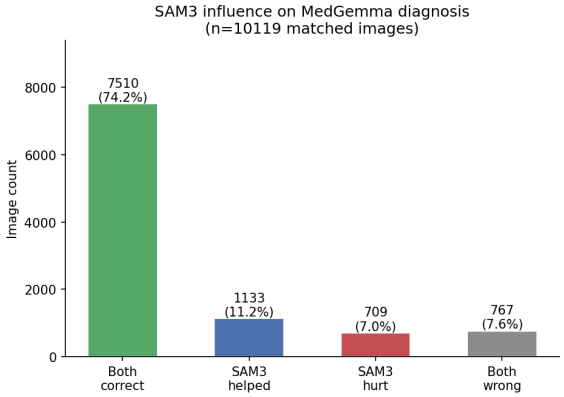


Fig. 5. Per-image outcome breakdown across $n = 10{,}119$ matched images. SAM3 helped: standalone MedGemma was wrong but the pipeline was correct. SAM3 hurt: standalone was correct but the bounding box misled the pipeline into an error.

(7.6%), representing cases that are difficult regardless of spatial attention.

The 709 hurt cases break down asymmetrically by ground-truth class. Among 577 hurt normal images, the dominant failure is SAM3 producing a false-positive region that causes MedGemma to predict tumor (513 cases; 88.9%), with a smaller number predicting other abnormalities (55) or multiple sclerosis (8). This is a direct consequence of SAM3 generating false-positive detections on normal anatomy—vessels, sulci, and choroid plexus—and MedGemma treating the bounding box as confirmatory evidence. Among the 132 hurt tumor images, most are redirected to other abnormalities (77) or normal (32), with the remainder scattered across cyst (11), multiple sclerosis (8), and stroke (4). The most affected subtype is the generic "tumor" class (84 hurt cases), followed by schwannoma (19) and glioma (12), suggesting that tumors with subtle or atypical enhancement patterns are most vulnerable to bounding box mislocalisation.

The 1,133 helped cases represent images where MedGemma's standalone prediction was wrong but the bounding box provided sufficient spatial context to recover the correct label. These are predominantly tumor images where standalone MedGemma had predicted either normal (486 cases) or multiple sclerosis (329 cases). Together, the hurt and helped analyses confirm that SAM3's role in the pipeline is not uniformly beneficial or harmful: it functions as a strong spatial cue that amplifies MedGemma's confidence in whatever region it is shown. When the bounding box is accurate, this is corrective; when it is a false positive, it overrides an otherwise correct prediction. This motivates a targeted intervention: a mask-fraction or confidence threshold on SAM3's output could suppress the low-coverage false positives that are disproportionately responsible for the normal-scan specificity collapse, while preserving beneficial spatial guidance on true tumor cases.

VII. Discussion

A. SAM3 Limitations for MRI

SAM3's limited zero-shot performance (Dice $= 0.189$, pixel-level) reflects fundamental domain mismatch rooted in the specific characteristics of brain tumor MRI. Natural-image segmentation relies on colour, high-contrast edges, and well-defined object boundaries that are largely absent in clinical neuroimaging. Brain tumors, particularly high-grade gliomas, exhibit infiltrative margins that blend gradually with surrounding oedematous tissue, leaving no sharp boundary for the model to detect. Signal intensity varies substantially across patients, tumor grades, and MRI sequences, so the same pathology appears differently across acquisitions. Acquisition-protocol variation—field strength, slice thickness, contrast-agent dose—introduces additional distributional shift that a model trained on web-scale natural images cannot anticipate. This is consistent with findings in the broader SAM adaptation literature [9]–[12]. The relatively high sensitivity (0.397) compared to Dice and IoU indicates that SAM3 does achieve partial overlap with tumor regions but generates excessively large masks, with pixel-level precision of only 0.097—meaning fewer than 1 in 10 predicted foreground pixels belongs to true tumor tissue.

Linear probing mitigates this without modifying the backbone. On BraTS the probe achieves Dice $= 0.836$, approaching the performance of several lightly supervised segmentation approaches despite using only a frozen backbone and a minimal linear head. This suggests that SAM3 features encode meaningful spatial representations that can be leveraged for medical segmentation with very limited task-specific learning.

The superiority of the linear probe over zero-shot is highly significant ($W = 7863$, $p = 1.98 \times 10^{-22}$, Wilcoxon signed-rank, $n = 125$ paired cases), ruling out sampling effects.

On MSLesSeg and ISLES 2022, however, the probe reaches only Dice $= 0.263$ and 0.408 respectively (Section VI-B). This gap reflects fundamental constraints that are more acute for MS and stroke than for brain tumors. MS white-matter plaques are typically 2–10 mm in diameter and often iso-intense with

surrounding white matter. SAM3 downsamples inputs to a coarse spatial feature grid (64×64), meaning that a 3 mm plaque may map to a single feature vector, destroying the spatial detail required for accurate delineation. The resulting $\sim$50:1 background-to-foreground ratio overwhelms a single linear layer even under inverse-frequency weighting. Stroke lesions are larger on average, but DWI acquisitions suffer from susceptibility artefacts and EPI distortion that further confound feature-level discrimination. Together, these domain-specific challenges—sub-millimetre lesion sizes, severe class imbalance, and acquisition artefacts absent from SAM3's training distribution—explain why the frozen backbone underperforms substantially on these tasks. Addressing them likely requires a shallow non-linear probe, focal or Dice-weighted loss, and substantially larger training cohorts—all while leaving the frozen backbone intact.

B. MedGemma Strengths and Weaknesses

Weak sequence and plane prediction likely reflects training distribution limits or prompt design: these attributes require low-level signal-intensity integration rather than semantic content recognition. The extreme glioma bias in subtype classification (69.8% vs. $\leq 5.3\%$ for all other classes) suggests a frequency prior from training data rather than discriminative morphological reasoning for rare types. The self-reported high diagnosis confidence ($\mu = 0.948$) does not necessarily imply calibration; formal ECE measurements are needed to assess whether the model's stated confidence aligns with empirical accuracy.

C. Spatial Attention and the False-Positive Problem

The SAM3 + MedGemma pipeline demonstrates a general principle: providing a VLM with an explicit spatial prior via a bounding box overlay can substantially improve classification when the upstream detector is accurate. By drawing the SAM3-detected region directly onto the full image, MedGemma retains full anatomical context while receiving a strong visual cue about where the pathology lies. However, false positives propagate directly as classification errors, and the pipeline also introduces a coarsening effect on fine-grained subtype predictions: the presence of a bounding box appears to bias MedGemma toward the generic "tumor" label rather than discriminating among subtypes.

D. Ethical Considerations and Human Oversight

Deploying AI models in clinical diagnostic workflows carries ethical obligations beyond technical performance. Three considerations are directly relevant to the models studied here.

Asymmetric error consequences. In the SAM3 + MedGemma pipeline, SAM3 false positives caused 25.5% of normal scans to be incorrectly flagged as tumor-positive. In a clinical context, a false-positive detection can trigger unnecessary follow-up imaging, invasive biopsy, or patient anxiety. Conversely, the 10.2% of tumor cases where SAM3 produced no mask represent potential delayed diagnoses. Any deployment must therefore be evaluated not only by aggregate accuracy but by the clinical cost of each error type.

Human oversight. Our results confirm that both SAM3 and MedGemma are insufficient for standalone clinical use. They are appropriately understood as tools to assist, not replace, radiologists. AI-generated outputs should be reviewed by a qualified clinician before influencing any diagnostic or treatment decision. Practical mechanisms for maintaining meaningful oversight include explicit uncertainty estimates, per-prediction confidence thresholds, and audit trails that preserve the original images alongside AI outputs.

Interpretability. SAM3 and MedGemma are largely opaque: neither segmentation masks nor classification responses come with explanations of what visual evidence drove the output. The bounding-box overlay improves spatial grounding but does not expose MedGemma's internal reasoning. Interpretability tools—attention visualisation, saliency maps, or concept-based explanations—are necessary before clinicians can reliably audit AI decisions in high-stakes settings such as neuro-oncology.

VIII. Conclusions

We have presented a three-part study examining SAM3 and MedGemma for brain tumor analysis. Key findings are:

- **SAM3 zero-shot on BraTS 2021:** Dice $= 0.189$, IoU $= 0.124$, Sensitivity $= 0.397$, Precision $= 0.097$ across all 125 cases—insufficient for standalone clinical use, but indicating partial localisation capacity.
- **Linear probe:** Frozen SAM3 features support effective segmentation learning with a minimal head (pixel-level Dice $= 0.836$, per-case mean $= 0.801$), a gain of $+0.647$ (pixel-level) Dice over zero-shot. Gains are smaller for MS and stroke segmentation, reflecting severe class imbalance and small training sets.
- **MedGemma standalone:** Strong on coarse tasks (modality: 90.5%, tumor/normal: 81.6%) but weak on fine-grained attributes (sequence: 24.5%, subtype: 30.7%) with a pronounced glioma bias. Self-reported confidence (0.948) requires formal calibration evaluation.
- **SAM3 + MedGemma pipeline:** Improves tumor detection (96.3%) and plane prediction (51.6%), but degrades normal-scan specificity (41.3%) due to false-positive region propagation, and collapses subtype accuracy (17.0%) via coarse-label drift.

Future work includes: a retrospective mask-fraction threshold sweep using logged SAM3 outputs to recover normal-scan specificity without re-running inference; training a normal/abnormal classifier between SAM3 and MedGemma; prompt engineering for rare subtype discrimination; extending to multi-class BraTS sub-region segmentation; and formal ECE-based calibration evaluation for MedGemma's self-reported confidence scores.

IX. Funding

Funded by the European Union under Horizon Europe project ChatMED - Bridging Research Institutions to Catalyze

Generative AI Adoption by the Health Sector in the Widening Countries (grant agreement ID: 101159214).

REFERENCES

[1] N. Ravi *et al.*, "SAM 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[2] A. Sellergren *et al.*, "MedGemma technical report," Google DeepMind, Tech. Rep., 2025. [Online]. Available: https://arxiv.org/abs/2507.05201

[3] Center for Biomedical Image Computing and Analytics (CBICA), "RSNA-ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge 2021," 2021, accessed: 2026-03-03. [Online]. Available: https://www.med.upenn.edu/cbica/brats2021/

[4] J. Cheng, "Brain tumor dataset," 2017. [Online]. Available: https://doi.org/10.6084/m9.figshare.1512427

[5] A. Hamada, "Br35H: Brain tumor detection 2020," 2020, accessed: 2026-01-11. [Online]. Available: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-743-detection/data

[6] F. Fernando, "Brain tumor MRI images 44 classes," 2022. [Online]. Available: https://www.kaggle.com/datasets/fernando2rad/brain-tumor-mri-images-44c

[7] ——, "Brain tumor MRI images 17 classes," 2022. [Online]. Available: https://www.kaggle.com/datasets/fernando2rad/brain-tumor-mri-images-17-classes

[8] A. Kirillov *et al.*, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 3992–4003.

[9] T. Chen *et al.*, "SAM fails to segment anything? — SAM-Adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, medical image segmentation, and more," *arXiv preprint arXiv:2304.09148*, 2023.

[10] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, Jan. 2024. [Online]. Available: http://dx.doi.org/10.1038/s41467-024-44824-z

[11] X. Huang *et al.*, "SAMed: On the transferability of segment anything model to medical images," *arXiv preprint arXiv:2304.02700*, 2023.

[12] H. Wang *et al.*, "SAM-Med3D: General-purpose segmentation models for volumetric medical images," *arXiv preprint arXiv:2310.15161*, 2024.

[13] X. Gao and K. Lu, "RefSAM3D: Adapting SAM with cross-modal reference for 3D medical image segmentation," *arXiv preprint*, 2024.

[14] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," *arXiv preprint arXiv:1610.01644*, 2016. [Online]. Available: https://arxiv.org/abs/1610.01644

[15] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Universeg: Universal medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 438–21 451.

[16] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.

[17] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.

[18] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, and O. Oktay, "Making the most of text semantics to improve biomedical vision–language processing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[19] M. Moor, Q. Huang, S. Wu, M. Yasunaga, C. Zakka, Y. Dalmia, E. P. Reis, P. Rajpurkar, and J. Leskovec, "Med-flamingo: a multimodal medical few-shot learner," in *Proceedings of Machine Learning for Health (ML4H)*, 2023.

[20] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.

[21] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and explainable region-guided radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[22] W. O. Ikezogwo, K. Zhang, M. S. Seyfioglu, F. Ghezloo, L. Shapiro, and R. Krishna, "Medicalnarratives: Connecting medical vision and language with localized narratives," *arXiv preprint arXiv:2501.04184*, 2025.

[23] S. Kyung, J. Seo, H. Lim, D. Kim, H. Park, J. Sung, J. Kim, W. Jo, Y. Nam, and N. Kim, "Medregion-ct: Region-focused multimodal llm for comprehensive 3d ct report generation," *arXiv preprint arXiv:2506.23102*, 2025.

[24] K. Jin, Q. Sun, D. Kang, Z. Luo, T. Yu, W. Han, Y. Zhang, M. Wang, D. Shi, and A. Grzybowski, "Grounded report generation for enhancing ophthalmic ultrasound interpretation using vision-language segmentation models," *npj Digital Medicine*, 2026.

[25] K. Sun, S. Xue, F. Sun, H. Sun, Y. Luo, L. Wang, S. Wang, N. Guo, L. Liu, T. Zhao *et al.*, "Medical multimodal foundation models in clinical diagnosis and treatment: Applications, challenges, and future directions," *Artificial Intelligence in Medicine*, p. 103265, 2025.

[26] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[27] C. Ludwig, K. Namdar, and F. Khalvati, "AI-driven MRI-based brain tumour segmentation benchmarking," *arXiv preprint arXiv:2506.20786*, 2025.

[28] S. Bakas *et al.*, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, p. 170117, 2017.

[29] Center for Biomedical Image Computing and Analytics (CBICA), "Multimodal Brain Tumor Segmentation Challenge 2020: Data," 2020, accessed: 2026-03-03. [Online]. Available: https://www.med.upenn.edu/cbica/brats2020/data.html

[30] BraTS Organizers, "BraTS Name Mapping CSV: Cross-Year Subject Mapping BraTS'17–'20," 2020, official Synapse file from syn25829067; confirms TCGA/CBICA/TMC only. [Online]. Available: https://www.synapse.org/Synapse:syn25829067

[31] The Cancer Imaging Archive (TCIA), "UCSF-PDGM Pre-operative Diffuse Glioma MRI Collection," 2021, accessed: 2026-03-03. [Online]. Available: https://www.cancerimagingarchive.net/collection/ucsf-pdgm/

[32] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927. [Online]. Available: http://www.jstor.org/stable/2276774

[33] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, p. 153–157, 1947.

[34] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *CoRR*, vol. abs/1706.04599, 2017. [Online]. Available: http://arxiv.org/abs/1706.04599

[35] A. Malara *et al.*, "MSLesSeg: A benchmark dataset for multiple sclerosis lesion segmentation," in *Proceedings of the MICCAI Challenge on Multiple Sclerosis Lesion Segmentation*, 2023.

[36] M. R. Hernandez Petzsche *et al.*, "Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset," *Scientific Data*, vol. 9, no. 1, p. 762, 2022.