BTReport: A Framework for Brain Tumor Radiology Report Generation with Clinically Relevant Features

Juampablo E. Heras Rivera[*1] JEHR@UW.EDU
[1] *University of Washington*

Dickson T. Chen[*1] DTCHEN19@UW.EDU
Tianyi Ren[1] TR1@UW.EDU
Daniel K. Low[2] DALOW@UW.EDU
[2] *University of Washington School of Medicine*

Jacob Ruzevick[2] RUZEVICK@UW.EDU
Asma Ben Abacha[3] ABENABACHA@MICROSOFT.COM
[3] *Microsoft Health AI*

Alberto Santamaria-Pang[3] ALBERTO.SANTAMARIAPANG@MICROSOFT.COM
Mehmet Kurt[1] MKURT@UW.EDU

Editors: Under Review for MIDL 2026

Abstract

Recent advances in radiology report generation (RRG) have been driven by large paired image-text datasets; however, progress in neuro-oncology RRG has been limited due to a lack of open paired image-report datasets. Here, we introduce BTReport, an open-source framework for brain tumor RRG that constructs natural language radiology reports using reliably extracted quantitative imaging features. Unlike existing approaches that rely on large general-purpose or fine-tuned vision-language models for both image interpretation and report composition, BTReport first performs deterministic feature extraction of clinically-relevant features, then uses large language models only for syntactic structuring and narrative formatting. By separating RRG into a deterministic feature extraction step and a report generation step, the generated reports are completely interpretable and less prone to hallucinations. We show that the features used for report generation are predictive of key clinical outcomes, including survival time, and reports generated by BTReport are more closely aligned with reference clinical reports than existing baselines for RRG. Finally, to further research in neuro-oncology RRG, we introduce BTReport-BraTS, a companion dataset that augments BraTS imaging with synthetically generated radiology reports produced with BTReport. Code for this project can be found at: https://github.com/KurtLabUW/BTReport.

Keywords: Brain MRI, Radiology report generation, VASARI, Midline shift, Open dataset, Multimodal learning, Neurooncology

[*] Contributed equally

1. Introduction

Radiology is a medical specialty that employs a variety of imaging modalities (e.g., x-ray, computed tomography (CT), multi-parametric magnetic resonance imaging (mpMRI)) for the detection and monitoring of human disease. The radiology report contains a detailed summary of imaging findings, providing insights into a patient's condition crucial for diagnosis and clinical decision making. With a growing aging population, the demand for radiology services is expected to increase between 16.9% to 26.9% by 2055, while attrition in radiology also continues to increase (Christensen et al., 2025). As the gap between physician workload and available workforce widens, assisted radiology report generation (RRG) is positioned to help address this unmet clinical need.

RRG leverages advances in artificial intelligence (AI) to extract quantitative imaging markers from raw unstructured data in an automated manner. In clinical workflows, RRG promises to improve data quality, repeatability, factual completeness, and timeliness of radiology reporting. The adoption of vision-language models (VLMs) in RRG has led to substantial advances, allowing models to jointly reason over medical images and textual findings (Liu et al., 2019; Chen et al., 2020; Sellergren et al., 2025). These developments have been primarily driven by large-scale chest x-ray datasets such as MIMIC-CXR (Johnson et al., 2019) and IU X-ray (Demner-Fushman et al., 2016), which provide over 300,000 chest x-ray images paired with clinical reports. However, accessible image-report datasets across other radiology specialties are limited. For example, in neuro-oncology, large neuroimaging datasets are made openly available through efforts such as BraTS (de Verdier et al., 2024), but paired text reports for VLM training are lacking.

Here we introduce BTReport, a two-stage framework for brain tumor RRG which first deterministically extracts clinically-relevant imaging features, including patient metadata, VASARI features, and automated 3D midline shift measurement from mpMRI. These features are then ingested by large language models (LLMs) for clinical reasoning and synthetic report generation. By grounding RRG with clinically-relevant imaging features extracted deterministically, our approach reduces hallucinations, and lowers the likelihood of critical detail omission, a key challenge with LLM-based RRG (Wu et al., 2025).

Our contributions are as follows: **a)** a scalable framework for automated report generation for brain tumors using deterministic neuroimaging features (Section 5.2), **b)** a robust and interpretable 3D midline shift (MLS) estimation algorithm (Section 5.2.2), **c)** demonstration of clinical feature relevance through retrospective predictive modeling of overall survival (Section 5.1.2), and **d)** release of BTReport-BraTS, an open-source companion dataset integrating anatomical and pathological descriptors (Section 4.2.1).

2. Related Work

A variety of approaches have been proposed for the task of image-paired RRG and generally fall into one- and two-stage frameworks. The leading paradigm for report generation involves training monolithic VLM foundation models to extract image features and generate reports in one step, such as MedGemma (Sellergren et al., 2025) and MedPaLM-2 (Singhal et al., 2023). Approaches in neuro-oncology following this paradigm include *TextBraTS* (Shi et al., 2025), which directly prompts GPT-4 models (Achiam et al., 2023) with videos of 2D mpMRI axial slices and corresponding tumor segmentation masks to extract quantitative features for RRG. While reports generated with this approach improved segmentation performance, it relies on a general-purpose VLM for image interpretation, and is thus prone to hallucinations from out-of-domain inference. Furthermore, the vision encoder of GPT-4 models is not optimized for quantitative measurements such as MLS estimation.

Unlike the one-stage frameworks described previously, approaches such as *From Segmentation to Explanation* (Valerio et al., 2025) adopt a two-stage approach: first, tumor location is derived from tumor-brain-atlas co-registration, then this is provided to a LLM for RRG. However, these explanations are not focused on clinically-relevant features, and reports only contain information about the tumor location. AutoRG-Brain (Lei et al., 2024) also uses a two-stage framework. In the first stage, tumor localization and region of interest (ROI) selection is done using anatomical segmentations. In the second stage, a fine-tuned VLM, which takes ROI image features and text prompts as inputs, is used for RRG. This approach uses deterministic features to focus the VLM generation on relevant regions of the image. However, AutoRG-Brain relies on a vision encoder for quantitative measurements (which is prone to hallucination) and the reports used for fine-tuning have limited clinically-relevant features. The framework proposed in *RadGPT* (Bassi et al., 2025) leverages a deterministic feature extraction step guided by clinical relevant features from abdominal CTs, then uses LLMs for syntactic structuring. This approach produces radiology reports that more closely resemble reference reports in comparison to end-to-end report generation models. However, it remains unclear whether *RadGPT* can be applied for neuro-oncology RRG from mpMRI.

Our approach is a two-stage framework with an extensive quantitative feature extraction step closest to that in *RadGPT*. In contrast to previous approaches that rely on VLMs for image analysis, our approach significantly reduces hallucinations by leveraging validated open-source algorithms for quantitative feature extraction (midline shift calculation, brain tumor statistics) and tumor localization. These quantitative features are then fed into a general-purpose LLM, leveraging its reasoning capabilities for RRG.

3. Clinical Background

Brain tumors represent a diverse group of central nervous system (CNS) malignancies, with gliomas comprising roughly 80% of all malignant primary brain tumors (Ostrom et al., 2014). Glioblastoma multiforme (GBM) is the most aggressive subtype, with an incidence of approximately 3.2 cases per 100,000 adults (Thakkar et al., 2014) and a median survival of 9-16 months following initial diagnosis (Bi and Beroukhim, 2014). Clinical evaluation typically begins with the presentation of neurological symptoms, prompting acquisition of mpMRI for radiological assessment. Radiographic findings guide decisions regarding surgical resection and biopsy, underscoring the need for robust MRI-based tumor characterization.

Advances in computer vision have automated GBM segmentation, allowing for detailed characterizations of tumor structure from mpMRI, such as the contrast-enhancing lesion, the necrotic core, and surrounding peritumoral edema, with strong agreement with manual annotations from expert raters (Menze et al., 2014). While these breakthroughs have been essential for quantifying tumor morphology, segmentations alone do not contain the clinical context required to understand the effects of the tumor on the brain. Instead, neuroradiologists assess additional imaging-derived features, including invasion of critical brain compartments (e.g., white matter, gray matter, vasculature, and basal cisterns) and tumor mass effect quantified by MLS measurement. This anatomical context provides a richer description of GBM, paving the path for improved risk stratification, personalized treatment plans, and survival outcome forecasting. Here, we incorporate anatomical context by using imaging features commonly reported by radiologists for improved RRG.

4. Data

4.1. Existing datasets

4.1.1. University of Washington Dataset (UWMC)

Pre-operative mpMRI scans and radiology reports were collected from a retrospective cohort of GBM patients (n=85) treated at the University of Washington Medical Center (UWMC). The following inclusion criteria were used: 1) confirmed histopathologic diagnosis of GBM; 2) availability of pre-operative mpMRI imaging including T1, T2, T1c, and T2-FLAIR; and 3) corresponding pre-operative diagnostic radiology reports. Imaging volumes were pre-processed using CaPTk (Davatzikos et al., 2018; Pati et al., 2020), following the same steps as BraTS 2017-2023, including: DICOM to NIfTI conversion, SRI24 co-registration, 1 mm isotropic resampling, skull-stripping, and tumor segmentation generated using the DeepMedic (Kamnitsas et al., 2017) deep-learning model. The accompanying pre-operative radiology reports were written by fellowship-trained radiologists and selected as the clinical reference standard. The final dataset contains fully de-identified MRI volumes, de-identified associated radiologist-authored reports, and tumor segmentation masks.

4.1.2. Survival Analysis Dataset

From the BraTS'23 dataset (Adewole et al., 2023), a smaller cohort of mpMRI cases (n=461) were used for survival analyses. Cases were selected based on the availability of five minimum metadata entries: (1) age at initial diagnosis, (2) biological sex, (3) confirmed methylation status of O6-methylguanine-DNA methyltransferase (*MGMT*), (4) mutation status of isocitrate dehydrogenase 1 (*IDH1*), and (5) overall survival or equivalent survival quantification representing the number of days between radiological diagnosis and reported days to known death. We collected these demographic and genomic features from multiple publicly available collections of GBM cases, including those from the University of California San Francisco (UCSF-PDGM)(Calabrese et al., 2022), University of Pennsylvania (UPenn-GBM) (Bakas et al., 2021), Clinical Proteomic Tumor Analysis Consortium (CPTAC-GBM) (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC), 2018), and Cancer Genome Atlas from the Cancer Imaging Archive (TCGA-LGG & TGCA-GBM) (Pedano et al., 2016) (Scarpace et al., 2016).

4.2. Novel dataset

4.2.1. BTReport-BraTS: A Companion Dataset for BraTS RRG

Pre-operative mpMRI cases from the combined training and validation splits of the BraTS 2023 Adult Glioma (BraTS'23) dataset (n=1,470 cases) were used to develop the BTReport-BraTS dataset, an open-source companion dataset generated using the BTReport framework described below (Section 5.2). For each case, corresponding midline segmentations, extracted metadata, structured summary reports, and radiology reports generated using the BTReport framework 5.2 are included and openly available.

5. Methods

5.1. Feature Selection

5.1.1. Semantic Clustering of Common Radiology Findings

To understand the most frequently observed concepts in real radiology reports, we extract all structured "fact" statements generated by the TBFact (Blondeel et al., 2025) metric across subjects in the UWMC dataset. All sentences across subjects were then embedded using a lightweight sentence transformer (all-MiniLM-L6-v2), producing a dense semantic vector representation. We then applied hierarchical agglomerative clustering (cosine distance, average linkage) with no predefined cluster count, allowing coherent groups of related statements to emerge from the data. Finally, once data was clustered, the Gemma 3 27B (Gemma Team et al., 2025) pre-trained LLM was used to summarize examples from each cluster into a short 3–8 word cluster description. The resulting set of cluster descriptions and frequencies provides an interpretable view of the themes and facts presented in generated reports, informing the quantitative features deterministically extracted in BTReport. The top 35 most frequent clusters found using this approach and example sentences for each cluster can be found in Appendix A.

5.1.2. Survival Modeling

To assess whether the extracted features were predictive of key clinical outcomes, we evaluated their association with overall survival using Kaplan–Meier survival analysis implemented through the `lifelines` (Davidson-Pilon, 2024) library. Differences between survival curves were assessed using the log-rank test. To quantify relative risk, we fit a Cox proportional hazards model and reported hazard ratios with 95% confidence intervals. Both the log-rank test and Cox model p-values were used to determine statistical significance. Kaplan-Meier plots for all BTReport-extracted features are shown in (Appendix B).

5.2. BTReport Framework

For a given subject, the BTReport framework (Figure 1) consists of four sequential stages:

1. As outlined in Section 5.2.1, the subject's T1-weighted scan and tumor segmentation are used to generate tumor-robust anatomical segmentations, from which region-wise volumetric statistics are computed.
2. Midline-shift features are derived by propagating the hand-annotated MNI152 midline into subject space using the MNI152-to-subject deformation field (Section 5.2.2).
3. A subset of VASARI descriptors are extracted using a modified version of VASARI-auto (Ruffle et al., 2024), which operates on all segmentations obtained in the previous steps (Section 5.2.3).
4. Finally, a structured radiology report is generated from the clinically selected features (Section 5.2.4) using a large language model, producing narrative findings that aim to emulate expert clinical reporting (Section 5.2.5).

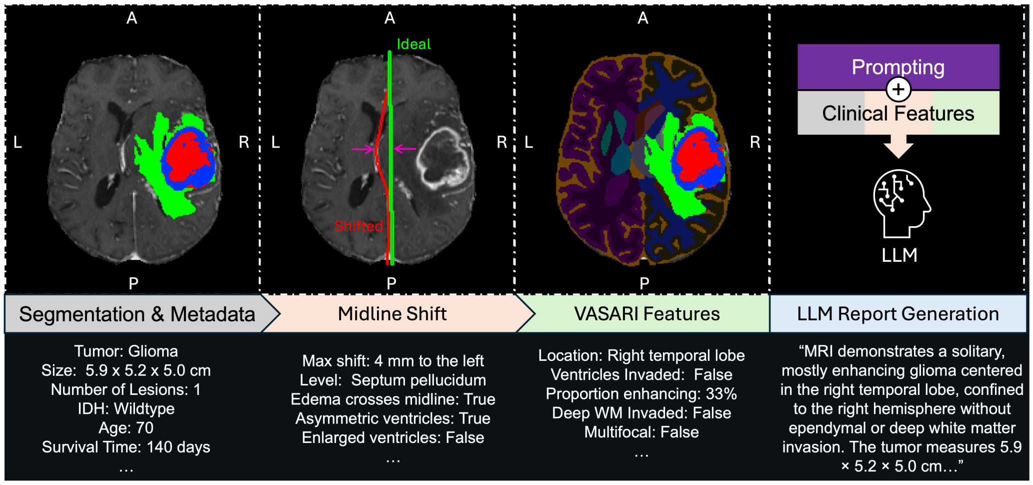
A
L
R
P
A
Ideal
L
R
Shifted
P
A
L
R
P
Prompting
+
Clinical Features
LLM
Segmentation & Metadata
Midline Shift
VASARI Features
LLM Report Generation
Tumor: Glioma
Size: 5.9 x 5.2 x 5.0 cm
Number of Lesions: 1
IDH: Wildtype
Age: 70
Survival Time: 140 days
…
Max shift: 4 mm to the left
Level: Septum pellucidum
Edema crosses midline: True
Asymmetric ventricles: True
Enlarged ventricles: False
…
Location: Right temporal lobe
Ventricles Invaded: False
Proportion enhancing: 33%
Deep WM Invaded: False
Multifocal: False
…
“MRI demonstrates a solitary, mostly enhancing glioma centered in the right temporal lobe, confined to the right hemisphere without ependymal or deep white matter invasion. The tumor measures 5.9 × 5.2 × 5.0 cm…”

Figure 1: BTReport Overview: First, a set of interpretable, clinically meaningful variables are deterministically extracted from each patient's imaging and metadata, including patient demographics, VASARI features, and 3D midline shift measurements. These features are then used to prompt an LLM to generate clinically-grounded radiology reports.

5.2.1. Segmentation Statistics

For a given subject, their 3D T1 scan and their corresponding tumor segmentation masks are inputted, with necrotic core, edema, and enhancing tumor sub-regions labeled according to the BraTS convention. These tumor segmentations can be obtained with radiologist manual annotation or an automatic segmentation algorithm. Next, the MNI152 atlas (Collins et al., 1999; Fonov et al., 2011, 2009) is registered into subject space and the subject scan is registered into the MNI152 space, producing two volumes and corresponding nonlinear registration transform maps. We refer to these as the MNI152-to-subject and subject-to-MNI152 registrations, respectively.

We then obtain anatomical segmentations by running SynthSeg (Billot et al., 2023) on the MNI152-to-subject volume. This substantially improves the robustness of SynthSeg in cases where the presence of tumor distorts tissue boundaries and automatic segmentation fails. We then merge the resulting anatomical labels with the tumor and midline segmentations to produce a unified mask containing both normal anatomy and pathological structures. Finally, we obtain features including tumor volume, number of lesions, proportion of tumor sub-regions, and the anatomical regions overlapping with the tumor.

5.2.2. 3D Midline Shift Estimation via Atlas-based Segmentation

Midline shift (MLS) is an intracranial pathology characterized by the displacement of brain tissue across the skull's midsagittal axis. MLS arises as a result of traumatic brain injury or tumor mass effects and is an indirect indicator of elevated intracerebral pressure. Estimation of MLS is done by identifying the axial slice with the largest deviation, as indicated by midline structures such as the septum pellucidum, the third ventricle, the fourth ventricle, or the falx cerebri. However, this estimation is subject to high inter-rater variability as there is not a standard procedure for axial slice level selection. Here, we propose a novel pipeline for MLS estimation based on clinical guidelines, using a deep learning atlas-based segmentation approach. Our approach leverages the robust registration capabilities of SynthMorph (Hoffmann et al., 2024) to register hand-annotated midline segmentations from a MNI152 atla template onto patient T1 scans. These are compared to an "ideal" midline defined by connecting the anterior and posterior points of the falx cerebri for each axial slice. By calculating the distance between the ideal and subject midlines at each voxel, we obtain highly-accurate 3D MLS estimations in seconds, giving a more complete picture in comparison to 2D automated or manual annotation methods. Furthermore, this approach has strong zero-shot generalization and can be applied to any MRI or CT scan.

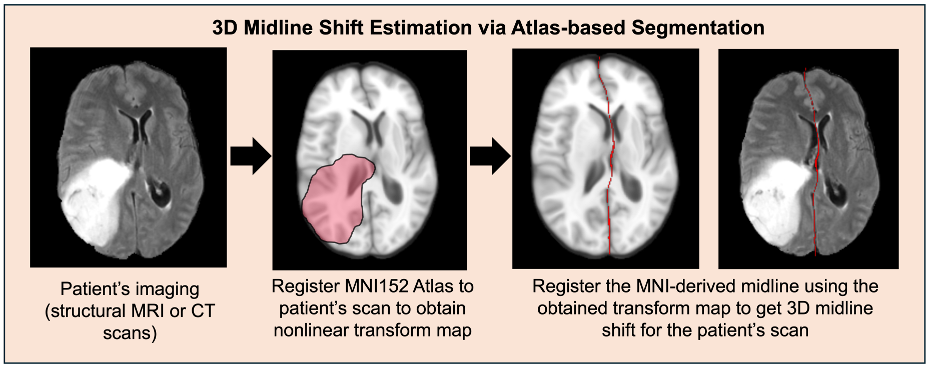
3D Midline Shift Estimation via Atlas-based Segmentation
Patient's imaging (structural MRI or CT scans)
Register MNI152 Atlas to patient's scan to obtain nonlinear transform map
Register the MNI-derived midline using the obtained transform map to get 3D midline shift for the patient's scan

Figure 2: 3D Midline Shift Estimation via Atlas-based Segmentation

5.2.3. VASARI Feature Extraction

To standardize neuroimaging-derived feature extraction and improve repeatability, tools such as VASARI (Visually AcceSAble Rembrandt Images) have been developed. The VASARI feature set quantitatively describes anatomical relationships between GBM and clinically relevant brain structures established in the literature. Furthermore, these relationships are also routinely included in neuroradiology reports and used by neurosurgeons to assess whether patients are candidates for surgical intervention. VASARI features have been used to accurately predict tumor histological grade (WHO I-IV grade), disease progression, molecular mutation status (e.g., *IDH1* WT/mutant, *MGMT* un/methylated), risk of recurrence, and overall patient survival (Jain et al., 2014; Nicolasjilwan et al., 2015; Peeken et al., 2019; Setyawan et al., 2024; Wang et al., 2021; Zhou et al., 2017). Here, we employ a modified variant of VASARI-auto (Ruffle et al., 2024), an automated labeling tool, which has been validated as non-inferior to radiologist VASARI annotations, and can be used to reduce inter-rater variance. The included variations make use of the subject-space anatomical segmentations, and midline segmentations extracted in Sections 5.2.1 and 5.2.2.

5.2.4. Selection of clinically relevant imaging features

Table 1 lists the quantitative imaging features used in BTReport for automated report generation. We assessed each feature using two criteria: 1) whether it could be directly used to obtain at least one of the top 35 concepts frequently reported in reference radiology reports from the UWMC dataset, and 2) verify if it could be directly used to obtain at least one of the top 35 concepts frequently reported in reference reports. Following the procedure outlined in Section 5.1.1 , each feature was assessed for whether it was a significant predictor of survival time according to the procedure outlined in Section 5.1.2.

Table 1: Summary of quantitative features used in BTReport. Feature groups are color-coded: gray = segmentation statistics, green = VASARI features, orange = midline features. Key: ● indicates the feature is among the top 30 most frequently reported concepts in real radiology reports, and ● indicates the feature is a significant survival predictor. Acronyms: WM-white matter; MLS-midline shift.

Total tumor volume (mL)	●	●	Ventricular Invasion	●	●	Level of max MLS	●	●
3D Lesion Sizes (cm)	●		Side of Tumor Epicenter	●		Max MLS (mm) + L/R	●	●
Proportion of Necrosis	●		Enhancement Quality	●		Edema crosses midline	●	●
Number of lesions	●		Enhancement thickness	●		ET Crosses midline	●	●
Proportion of Enhancing	●	●	Multiple satellites present	●		Asymmetrical Ventricles	●	
Proportion of Edema	●		Multifocal or Multicentric	●	●	Enlarged Ventricles	●	●
Cortical involvement	●		Deep WM invasion	●	●			
Tumor Location	●	●	Eloquent Brain Involved					

5.2.5. Synthetic Report Generation via LLMs

BTReport generates the *Findings* sections of radiology reports in a style which mimics that of a target institution through an in-context learning prompt (shown in Appendix C). The prompt contains examples of *Findings* sections from reference radiology reports from the target institution to provide style guidance. Additionally, the prompt includes detailed instructions to the LLM, instructing the model to only report facts directly derived from the provided reference reports. For RRG with BTReport, we experiment with two open-source pre-trained LLM models with reasoning capability: gpt-oss-120b (Agarwal et al., 2025) and Llama 3.1 70B Instruct (Grattafiori et al., 2024). Examples *Findings* sections from synthetically generated across these different frameworks can be found in Appendix D. All example reports used for prompting were de-identified and only local offline copies of LLMs were used for analysis to avoid cloud sharing of privileged medical data.

6. Evaluation Metrics

For evaluation, we used 30 paired image-text report datasets of GBM subjects in the UWMC cohort. We generated synthetic radiology reports from the images using multiple neuro-oncology RRG frameworks, including the BTReport pipeline, then compared generated report findings against real radiology reports using four evaluation metrics from RadEval (Xu et al., 2025), a unified open-source framework that evaluates radiology text based on:

- N-gram-based lexical similarity: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004)
- pre-trained contextualized embeddings: BERTScore (Zhang et al., 2020)
- clinically grounded scores: RaTEScore (Zhao et al., 2024)

Additionally, to assess the clinical correctness of content in generated reports, we used TBFact (Blondeel et al., 2025), an LLM-based factuality metric that evaluates generated reports based on three key criteria: factual inclusion, distortion, and omission.

7. Results

Here, we compare two BTReport variants (BTReport (gpt-oss:120B) and BTReport (LLaMA3:70B)) against existing neuro-oncology RRG frameworks, specifically AutoRG-Brain and Seg-to-Exp (described in Section 2). We evaluate lexical similarity and semantic/factual accuracy using the metrics described in Section 6. All systems are benchmarked on the same held-out dataset of 25 subjects under identical evaluation protocols to ensure consistent comparison. Paired significance testing was performed using Approximate Randomization (AR).

7.1. Lexical similarity of generated reports

Table 2: BLEU and ROUGE metrics. Best values per metric are highlighted in blue. †Following AR, both BTReport generated reports were superior to those generated by other frameworks across all evaluation metrics ($p < 0.0001$).

Framework	**BLEU-1**	**BLEU-2**	**ROUGE-1**	**ROUGE-2**
BTReport (gpt-oss:120B)†	0.260	0.139	0.390	0.120
BTReport (LLaMA3:70B)†	**0.265**	**0.151**	**0.393**	**0.128**
AutoRG-Brain (Lei et al., 2024)	0.165	0.0841	0.288	0.0784
Seg-to-Exp (Valerio et al., 2025)	0.0952	0.0387	0.178	0.0236

7.2. Factual accuracy of generated reports

Table 3: TBFact, BERTScore, and RaTEScore metrics. Best values per metric are highlighted in blue. † Following AR, both BTReport generated reports were superior to those generated by other frameworks across all evaluation metrics ($p < 0.0001$).

	TBFact (DeepSeek-R1)				**BERTScore**	**RaTEScore**
Framework	**Score**	**Prec.**	**Recall**	**F1**		
BTReport (gpt-oss:120B)†	**0.317**	**0.382**	**0.306**	**0.317**	**0.462**	**0.573**
BTReport (LLaMA3:70B)†	0.268	0.328	0.253	0.268	0.437	0.562
AutoRG-Brain (Lei et al., 2024)	0.211	0.311	0.194	0.211	0.347	0.492
Seg-to-Exp (Valerio et al., 2025)	0.105	0.133	0.145	0.107	0.166	0.414

8. Discussion

Tables 2 and 3 indicate that radiology reports generated using the BTReport framework demonstrate substantial performance increases over existing neuro-oncology RRG systems across both lexical and factual evaluation metrics. Both BTReport variants (BTReport (gpt-oss:120B) and BTReport (LLaMA3:70B)) outperformed AutoRG-Brain and Seg-to-Exp when evaluated using BLEU, ROUGE, BERTScore, RATEScore, and TBFact metrics. Further, improvements using the BTReport framework were statistically significant ($p < 0.0001$) when evaluated with Approximate Randomization. These results indicate that integrating reliably extracted quantitative features during yields synthetically generated reports that more closely align with radiologist clinical interpretations. Interestingly, the BTReport (LLaMA3:70B) variant outperformed the larger BTReport (gpt-oss:120B) variant based on lexical metrics, suggesting that LLaMa3 is able to more closely match the wording used in the reference radiology reports provided in the LLM prompt.

Overall, our findings support the two-stage report generation paradigm, suggesting that in medical imaging domains with limited data, adding quantitative features to prompts is an efficient way to generate reports and reduce the number of critical omissions. Furthermore, by using a two stage approach, it is possible for the end user to see which features were used to generate the final report, adding a layer of interpretability.

Our study had limitations which should be adressed in future work. While BTReport uses features that cover a range of the most reported quantitative features included in radiology reports, some still remain. These include things like white matter hyperintensity, basal cisterns, and stroke lesions. In future versions, these should be accounted for, as these are very common.

9. Conclusion

We present BTReport, a scalable two-stage framework for RRG grounded in quantitative clinically relevant imaging features extracted deterministically. We provide BTReport-BraTS, an open-source companion dataset containing accompanying anatomical features, metadata, and BTReport generated reports for mpMRI cases in the BraTS'23 dataset. By leveraging a two-stage framework, we increase the fidelity of synthetically generated reports to quantitative measurements, improve reliability of RRG. We hope the proposed framework is used to further research in VLM development in neuro-oncology applications.

Acknowledgments

The authors are grateful to Caitlin Neher, Leonardo Schettini, and Ankush Jindal for their helpful discussions. The authors are also grateful to James K. Ruffle, for making the VASARI-Auto implementation available.

10. Ethics

This study's activities were approved by the Institutional Review Board at the University of Washington (STUDY00022466). This research is in accordance with the principles embodied in the Declaration of Helsinki.

11. Funding Statement

The work of Juampablo Heras Rivera was partially supported by the U.S. Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0024386.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Maruf Adewole, Jeffrey D Rudie, Anu Gbdamosi, Oluyemisi Toyobo, Confidence Raymond, Dong Zhang, Olubukola Omidiji, Rachel Akinola, Mohammad Abba Suwaid, Adaobi Emegoakor, et al. The brain tumor segmentation (BraTS) challenge 2023: Glioma segmentation in sub-saharan Africa patient population (BraTS-Africa). *ArXiv*, pages arXiv–2305, 2023.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

S. Bakas, C. Sako, H. Akbari, M. Bilello, A. Sotiras, G. Shukla, J. D. Rudie, N. Flores Santamaria, A. Fathi Kazerooni, S. Pati, S. Rathore, E. Mamourian, S. M. Ha, W. Parker, J. Doshi, U. Baid, M. Bergman, Z. A. Binder, R. Verma, and C. Davatzikos. Multi-parametric magnetic resonance imaging (mpMRI) scans for de novo Glioblastoma (GBM) patients from the University of Pennsylvania Health System (UPENN-GBM). 2021. Dataset on The Cancer Imaging Archive (TCIA).

Pedro RAS Bassi, Mehmet Can Yavuz, Ibrahim Ethem Hamamci, Sezgin Er, Xiaoxi Chen, Wenxuan Li, Bjoern Menze, Sergio Decherchi, Andrea Cavalli, Kang Wang, et al. RadGPT: Constructing 3d image-text tumor datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23720–23730, 2025.

Wenya Linda Bi and Rameen Beroukhim. Beating the odds: extreme long-term survival with glioblastoma. *Neuro-Oncology*, 16(9):1159–1160, 09 2014. ISSN 1522-8517. doi: 10.1093/neuonc/nou166. URL https://doi.org/10.1093/neuonc/nou166.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023.

Matthias Blondeel, Noel Codella, Sam Preston, Hao Qiu, Leonardo Schettini, Frank Tuan, Wen-wai Yim, Smitha Saligrama, Mert Öz, Shrey Jain, et al. Healthcare Agent Orchestrator (HAO) for Patient Summarization in Molecular Tumor Boards. *arXiv preprint arXiv:2509.06602*, 2025.

E. Calabrese, J. Villanueva-Meyer, J. Rudie, A. Rauschecker, U. Baid, S. Bakas, S. Cha, J. Mongan, and C. Hess. The University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM). 2022. Dataset on The Cancer Imaging Archive (TCIA).

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

Eric W Christensen, Jay R Parikh, Alexandra R Drake, Eric M Rubin, and Elizabeth Y Rula. Projected US radiologist supply, 2025 to 2055. *Journal of the American College of Radiology*, 22(2):161–169, 2025.

D Louis Collins, Alex P Zijdenbos, Wim FC Baaré, and Alan C Evans. ANIMAL+ INSECT: improved cortical structure segmentation. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 210–223. Springer, 1999.

C. Davatzikos, S. Rathore, S. Bakas, S. Pati, M. Bergman, R. Kalarot, P. Sridharan, A. Gastounioti, N. Jahani, E. Cohen, H. Akbari, B. Tunc, J. Doshi, D. Parker, M. Hsieh, A. Sotiras, H. Li, Y. Ou, R. K. Doot, M. Bilello, Y. Fan, R. T. Shinohara, P. Yushkevich, R. Verma, and D. Kontos. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *Journal of Medical Imaging*, 5(1):011018, 2018. doi: 10.1117/1.JMI.5.1.011018.

Cameron Davidson-Pilon. lifelines: Survival Analysis in Python. 2024. Zenodo software release.

Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment MRI. *arXiv preprint arXiv:2405.18368*, 2024.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, March 2016. doi: 10.1093/jamia/ocv080.

Vladimir Fonov, Alan C Evans, Kelly Botteron, C Robert Almli, Robert C McKinstry, D Louis Collins, Brain Development Cooperative Group, et al. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327, 2011.

Vladimir S Fonov, Alan C Evans, Robert C McKinstry, CR Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, 2009.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, and Morgane Rivière. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The LLaMa 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Malte Hoffmann, Andrew Hoopes, Douglas N Greve, Bruce Fischl, and Adrian V Dalca. Anatomy-aware and acquisition-agnostic joint registration with SynthMorph. *Imaging Neuroscience*, 2:1–33, 2024.

Rajan Jain, Laila M Poisson, David Gutman, Lisa Scarpace, Scott N Hwang, Chad A Holder, Max Wintermark, Arvind Rao, Rivka R Colen, Justin Kirby, et al. Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor. *Radiology*, 272(2):484–493, 2014.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-Ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

Jiayu Lei, Xiaoman Zhang, Chaoyi Wu, Lisong Dai, Ya Zhang, Yanyong Zhang, Yanfeng Wang, Weidi Xie, and Yuehua Li. AutoRG-Brain: Grounded Report Generation for Brain MRI. *arXiv preprint arXiv:2407.16684*, 2024.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya 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, 2014.

National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The Clinical Proteomic Tumor Analysis Consortium Glioblastoma Multiforme Collection (CPTAC-GBM). 2018. Dataset on The Cancer Imaging Archive (TCIA).

Manal Nicolasjilwan, Ying Hu, Chunhua Yan, Daoud Meerzaman, Chad A Holder, David Gutman, Rajan Jain, Rivka Colen, Daniel L Rubin, Pascal O Zinn, et al. Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. *Journal of Neuroradiology*, 42(4):212–221, 2015.

Quinn T. Ostrom, Luc Bauchet, Faith G. Davis, Isabelle Deltour, James L. Fisher, Chelsea Eastman Langer, Melike Pekmezci, Judith A. Schwartzbaum, Michelle C. Turner, Kyle M. Walsh, Margaret R. Wrensch, and Jill S. Barnholtz-Sloan. The epidemiology of glioma in adults: a "state of the science" review. *Neuro-Oncology*, 16(7):896–913, 05 2014. ISSN 1522-8517. doi: 10.1093/neuonc/nou087. URL https://doi.org/10.1093/neuonc/nou087.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002. URL https://aclanthology.org/P02-1040/.

S. Pati, A. Singh, S. Rathore, A. Gastounioti, M. Bergman, P. Ngo, S. M. Ha, D. Bounias, J. Minock, G. Murphy, H. Li, A. Bhattarai, A. Wolf, P. Sridaran, R. Kalarot, H. Akbari, A. Sotiras, S. P. Thakur, R. Verma, R. T. Shinohara, P. Yushkevich, Y. Fan, D. Kontos, C. Davatzikos, and S. Bakas. The Cancer Imaging Phenomics Toolkit (CaPTk): Technical Overview. In *BrainLes 2019. Lecture Notes in Computer Science*, volume 11993, pages 380–394. Springer, 2020. doi: 10.1007/978-3-030-46643-5_38.

N. Pedano, A. E. Flanders, L. Scarpace, T. Mikkelsen, J. M. Eschbacher, B. Hermes, V. Sisneros, J. Barnholtz-Sloan, and Q. Ostrom. The Cancer Genome Atlas Low Grade Glioma Collection (TCGA-LGG). https://doi.org/10.7937/K9/TCIA.2016.L4LTD3TK, 2016. Dataset on The Cancer Imaging Archive (TCIA).

Jan C Peeken, Tatyana Goldberg, Thomas Pyka, Michael Bernhofer, Benedikt Wiestler, Kerstin A Kessel, Pouya D Tafti, Fridtjof Nüsslin, Andreas E Braun, Claus Zimmer, et al. Combining multimodal imaging and treatment features improves machine learning-based prognostic assessment in patients with glioblastoma multiforme. *Cancer medicine*, 8(1):128–136, 2019.

James K Ruffle, Samia Mohinta, Kelly Pegoretti Baruteau, Rebekah Rajiah, Faith Lee, Sebastian Brandner, Parashkev Nachev, and Harpreet Hyare. VASARI-auto: Equitable, efficient, and economical featurisation of glioma MRI. *NeuroImage: Clinical*, 44:103668, 2024.

L. Scarpace, T. Mikkelsen, S. Cha, S. Rao, S. Tekchandani, D. Gutman, J. H. Saltz, B. J. Erickson, N. Pedano, A. E. Flanders, J. Barnholtz-Sloan, Q. Ostrom, D. Barboriak, and L. J. Pierce. The Cancer Genome Atlas Glioblastoma Multiforme Collection (TCGA-GBM). https://doi.org/10.7937/K9/TCIA.2016.RNYFUYE9, 2016. Dataset on The Cancer Imaging Archive (TCIA).

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Nurhuda Hendra Setyawan, Lina Choridah, Hanung Adi Nugroho, Rusdy Ghazali Malueka, and Ery Kus Dwianingsih. Beyond invasive biopsies: using VASARI MRI features to predict grade and molecular parameters in gliomas. *Cancer Imaging*, 24(1):3, 2024.

Xiaoyu Shi, Rahul Kumar Jain, Yinhao Li, Ruibo Hou, Jingliang Cheng, Jie Bai, Guohua Zhao, Lanfen Lin, Rui Xu, and Yen-wei Chen. TextBraTS: Text-Guided Volumetric Brain Tumor Segmentation with Innovative Dataset Development and Fusion Module Exploration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 638–648. Springer, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Jigisha P. Thakkar, Therese A. Dolecek, Craig Horbinski, Quinn T. Ostrom, Donita D. Lightner, Jill S. Barnholtz-Sloan, and John L. Villano. Epidemiologic and Molecular Prognostic Review of Glioblastoma. *Cancer Epidemiology, Biomarkers & Prevention*, 23 (10):1985–1996, 09 2014. ISSN 1055-9965. doi: 10.1158/1055-9965.EPI-14-0275. URL https://doi.org/10.1158/1055-9965.EPI-14-0275.

Alberto G Valerio, Katya Trufanova, Salvatore de Benedictis, Gennaro Vessio, and Giovanna Castellano. From segmentation to explanation: Generating textual reports from MRI with LLMs. *Computer Methods and Programs in Biomedicine*, page 108922, 2025.

Jing Wang, Xiaoping Yi, Yan Fu, Peipei Pang, Huihuang Deng, Haiyun Tang, Zaide Han, Haiping Li, Jilin Nie, Guanghui Gong, et al. Preoperative magnetic resonance imaging radiomics for predicting early recurrence of glioblastoma. *Frontiers in Oncology*, 11: 769188, 2021.

David Wu, Fateme Nateghi Haredasht, Saloni Kumar Maharaj, Priyank Jain, Jessica Tran, Matthew Gwiazdon, Arjun Rustagi, Jenelle Jindal, Jacob M Koshy, Vinay Kadiyala, et al. First, do NOHARM: towards clinically safe large language models. *arXiv preprint arXiv:2512.01241*, 2025.

Justin Xu, Xi Zhang, Javid Abderezaei, Julie Bauml, Roger Boodoo, Fatemeh Haghighi, Ali Ganjizadeh, Eric Brattain, Dave Van Veen, Zaiqiao Meng, et al. RadEval: A framework for radiology text evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 546–557, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. RaTEScore: A Metric for Radiology Report Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, 2024.

Hao Zhou, Martin Vallières, Harrison X Bai, Chang Su, Haiyun Tang, Derek Oldridge, Zishu Zhang, Bo Xiao, Weihua Liao, Yongguang Tao, et al. MRI features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro-oncology*, 19(6):862–870, 2017.

Appendix A. Cluster analysis of common topics in reference reports

Table 4: Example sentences associated with the top 35 radiological concept clusters.

Cluster Description	Frequency	Two Example Sentences in this cluster
Lateral ventricle effacement/asymmetry; possible. . .	50	"The lateral ventricles are symmetric.", "The lateral ventricles are symmetric"
Midline shift presence and magnitude.. . .	27	"There is a 2 mm left midline shift", "There is a susceptibility millimeters rightward midline shift."
Cisterns appear normal, no obstruction.	19	"The basal cisterns are patent.", "The basal cisterns are patent"
Peritumoral/Vasogenic edema present. *(This. . .	18	"There is mild-to-moderate associated edema", "The first mass has associated surrounding vasogenic edema"
Lesion size and measurements reported.	16	"The lesion measures 4.0 x 3.5 cm", "The lesion measures 3.9 x 2.0 x 2.1 cm."
No acute intracranial hemorrhage or infarct.	15	"There is no acute infarct", "No acute infarct is seen"
Multiple intracranial masses present bilaterally.. . .	15	"There is a large irregular enhancing mass centered in the right frontal lobe", "There is a right frontal lobe mass"
Multiple, enhancing intracranial lesions present.	14	"The lesion originates in the anterior paramedian left frontal lobe", "There is a lesion in the right frontal lobe"
White matter disease, likely nonspecific etiology.	14	"There are scattered deep and periventricular white matter T2/FLAIR hyperintensities", "There is mild subcortical and periventricular white matter T2 FLAIR abnormality"
No mass effect/midline shift. (Alternatively: No. . .	13	"There is no shift in the brain", "There is no shift of the brain structures"
Ventricular system: normal or prominent.. . .	13	"The ventricles, sulci, and cisterns are normal", "The remaining ventricles, sulci, and cisterns are normal"
Restricted diffusion within the lesion(s).	9	"The solid components of the lesion demonstrate moderate diffusion restriction", "The lesion has peripheral areas of mild diffusion restriction"
Frontal/Temporal lobe FLAIR signal abnormality	9	"There is extensive T2/FLAIR signal abnormality in the right frontotemporal lobes", "There is surrounding FLAIR signal hyperintensity inferiorly going into the temporal lobe and posteriorly."
Herniation syndromes present on imaging. (or. . .	8	"There is suggestion of transtentorial herniation", "There is leftward subfalcine herniation"
Mass dimensions and size measurements.	8	"The mass measures 2.4 x 3.4 cm in axial cross-section", "The mass measures approximately 6.6 x 4.7 cm in transverse dimensions and 4.6 cm craniocaudally"
Uncal medialization, potentially impacting. . .	8	"There is right uncal medialization", "There is medialization of the right uncus"
No acute intracranial hemorrhage present.. . .	7	"No parenchymal hemorrhage is present", "There is no associated hemorrhage"
Mass effect on lateral ventricles.	7	"The mass extends along the ependymal surface of the right lateral ventricle", "The mass extends into the posterior horn of the right lateral ventricle"
Edema predominantly affecting frontal & temporal. . .	6	"There is perilesional edema along the predominantly anterior aspect of the medial frontal lobes", "The vasogenic edema extends to the frontal lobe"
Mass size and measurements. (Alternatively: Lesion. . .	6	"The mass measures 40 x 53 mm", "The mass measures approximately 58 x 44 x 44 mm"
Corpus callosum lesion, midline crossing/spread.	6	"The lesion extends into the splenium of the corpus callosum, crossing midline to the right", "There is ependymal spread along the body of the corpus callosum"
Diffusion restriction presence/absence &. . .	6	"There is no associated restricted diffusion", "There are areas of internal diffusion restriction and susceptibility"
Frontal horn effacement & ventricular asymmetry.	6	"There is partial effacement of the right frontal horn", "There are areas of subtle ependymal enhancement in the bilateral frontal horns"
Hemorrhagic lesion with restricted diffusion.	5	"The lesion restricts diffusion and has intralesional hemorrhage", "The mass is T2 hyperintense, contains multiple foci of internal hemorrhage, and demonstrates mottled diffusion restriction consistent with hypercellularity and/or necrosis."
Cistern effacement suggests mass effect. (Or, more. . .	5	"There is effacement of the right crural cistern", "There is partial effacement of the basal cisterns"
Sulcal effacement, widespread cortical. . .	5	"There is sulcal effacement involving the right parietal, posterior temporal, and occipital lobes", "There is mild sulcal effacement of the left occipital lobe."
Mass shows restricted diffusion on imaging. (Or,. . .	5	"There are patchy foci of restricted diffusion within the mass", "The second mass has diffusion restriction"
Corpus callosum mass/involvement. (or simply:. . .	5	"The mass extends into the right-sided genu of the corpus callosum", "The mass extends along the splenium of the corpus callosum"
Basal ganglia involvement with signal abnormality.	5	"The signal abnormality extends into the right basal ganglia, right thalamus, right cerebral peduncle, and right midbrain", "The hyperintensity involves the bilateral basal ganglia, with greater involvement on the left"
Ventricular size and morphology assessment.	4	"The third ventricle is near slitlike", "There is complete effacement of the third ventricle"
Midline shift at foramen of Monro	4	"There is a negative millimeters leftward midline shift at the level of the foramen of Monroe", "There is some millimeters of rightward midline shift at the level of the foramen of Monro"
No acute hydrocephalus present.	4	"There is no evidence of acute hydrocephalus", "There is no hydrocephalus"
Corpus callosum FLAIR edema/hyperintensity. . .	4	"The surrounding T2/FLAIR signal is similar and extends to the left splenium, septum pellucidum, and superior corpus callosum", "There is subtle patchy T2 FLAIR hyperintensity along the right body of the corpus callosum"
Peripheral mass enhancement characteristics. (Or. . .	4	"The first mass has peripheral enhancement with a nodular solid enhancing component", "The mass has irregular somewhat nodular peripheral enhancement"
Right cerebral peduncle lesion/mass effect.	4	"There is mass effect on the right cerebral peduncle", "The lesion has questionable extension into the posterior right cerebral peduncle"

Appendix B. Kaplan-Meier survival analysis plots of extracted features

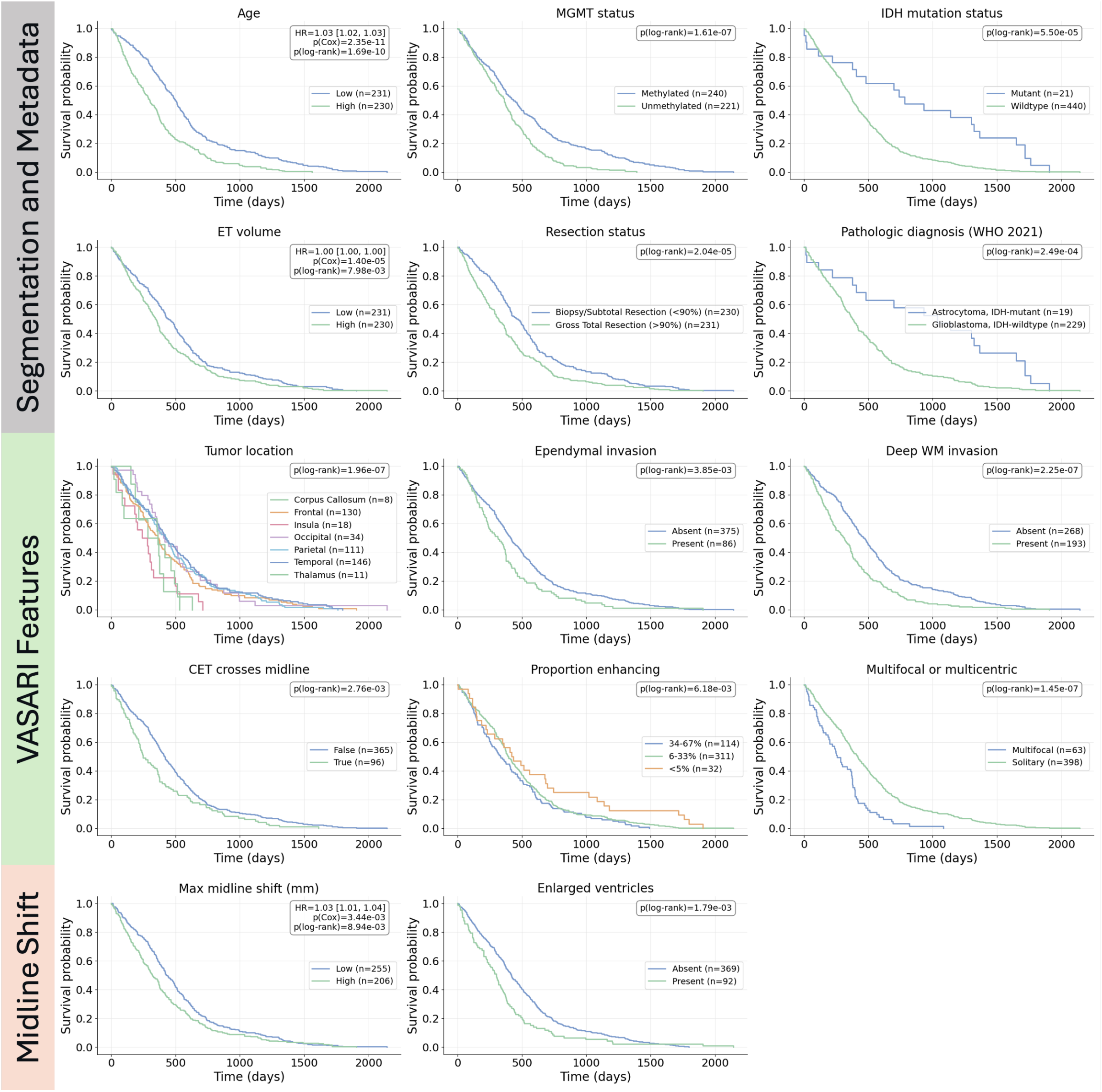

Figure 3: BTReport extracts a set of interpretable, clinically meaningful variables from each case, including patient demographics, VASARI features, and 3D midline shift measurements. These features summarize key aspects of tumor biology and mass effect that are routinely described radiology reports and neuro-oncology decision-making. Kaplan–Meier analyses show that many of these features are predictive of overall survival, highlighting their clinical relevance and motivating their use as structured inputs for radiology report generation.

Appendix C. LLM prompts used for report generation

BTReport Prompt

You are a radiologist generating a synthetic clinical MRI report.
Below are example FINDINGS sections taken from real brain tumor reports:

EXAMPLE FINDINGS:
`{example_findings}`

Your job is to generate a FINDINGS section in the same clinical style, but only using the METADATA provided below.
Please abide strictly to the following rules (follow them exactly).

1. Use only the metadata provided for quantitative statements. Do NOT hallucinate any information that is not directly inferable.
2. Include 10–20 clinically meaningful findings summarized in an anatomically descriptive manner. Prioritize describing abnormal or clinically significant observations.
3. Preserve the subsection structure from the example reports. Make sure to include the following subsections: MASS EFFECT & VENTRICLES and BRAIN / ENHANCEMENT.
4. Never mention imaging sequences other than T1n, T2w, T2 FLAIR, or T1-Gd. Do not mention diffusion, perfusion, spectroscopy, MRA, or other modalities (unless stated explicitly in metadata).
5. Do not mention structures or measurements not present in the metadata.
6. Mandatory Considerations: Make sure to include the following findings below if present in metadata. Remember to follow the sentence structure in the example reports.

    a) Maximum midline shift represented in mm units.

        i) Make sure to describe the magnitude and the direction of the midline shift.

        ii) Describe the anatomical level at which the (e.g., foramen of Monro, third and fourth ventricles, septum pellucidum).

        iii) If shift is minimal (e.g., < 5 mm), explicitly state the measurement as no shift, but still provide the measurement.

    b) If tumor mass effect is present, describe the mass effect on ventricles or surrounding brain structures.

        i) Include a description of ventricular effacement (if present), including which horn (anterior/posterior horn), and on which hemisphere it is observed.

    c) Comment on ventricular status. If effacement is present, describe the extent of the effacement or asymmetry. If ventricles are normal, explicitly state so (mirroring example reports). Use the following metadata fields in your description: "Asymmetrical Ventricles", "Enlarged Ventricles".

    d) Describe the size of the primary lesion, as well as any smaller secondary lesions (if present) represented in cm units. Use the 3D measurements from the metadata. Make sure to include the following:

        i) If multiple lesions exist, summarize number, dominant lesion, and laterality. Use the following metadata fields in your description: "Number of lesions" and "Multifocal or multicentric."

        ii) Anatomical location of lesion(s).

        iii) Use the following metadata fields in your description: "Tumor Location", "Side of Tumor Epicenter", and "Region Proportions."

    e) Describe enhancing characteristics. Use the following metadata fields in your description: "Enhancement quality", "Thickness of enhancing margin", "Proportion Enhancing".

        i) Describe enhancement style (e.g., rim-enhancing, mildly enhancing, peripheral enhancing, multilobular enhancing) only if explicitly supported.

        ii) Describe edematous tissue. Use the following metadata fields in your description: "ED volume", Whether edema crosses midline, "Proportion of edema".

        iii) Describe vasogenic edema and its extent only if metadata supports it.

    f) Describe invasion and involvement. Use the following metadata fields in your description: "Cortical involvement", "Deep WM invasion", "Ependymal invasion", "Eloquent Brain Involvement".

    g) Describe the necrosis if present. Use the metadata field "Proportion Necrosis" to describe the central foci of necrosis.

METADATA (for subject `{subject_id}`):
`{metadata_json}`

Write the **FINDINGS** section now, using clinical radiology language.

Appendix D. Example generated reports

The following key represents key clinically-relevant anatomical concepts describing GBM tumor location in relation to the brain. The following color scheme compares model-generated findings to ground truth findings based on:

- Green highlight indicates statements that match the reference (fully correct).
- Yellow highlight indicates statements that are partially correct or contextually close.
- Red highlight indicates statements that contradict the reference or contain factual errors.
- Gray highlight indicates statements in the ground truth that were referenced in reports frequently generated by other RRG methods.

****** Specific mpMRI imaging modality unavailable.
//

Model comparison of RRG output: Findings Section

Ground Truth Findings

MASS EFFECT & VENTRICLES: Prominent leftward midline shift by approximately 14 mm (XXX/XXX). There is also medialization of right uncus. The basal cisterns are partially effaced.

BRAIN/ENHANCEMENT: A large irregular enhancing lesion centered within the right temporal lobe with significant mass effect. Restricted diffusion****** is noted within the enhancing portion of the lesion, which contains susceptibility artifact suggestive of microhemorrhages/angioinvasion.****** The lesion measures approximately 6.8 x 4.4 x 4.8 cm (AP, TV, CC). There is probable small subependymal enhancement at the atria of right lateral ventricle (XXX/XXX, XXX/XXX). No acute hematoma or infarct is seen.

Generated Findings

Framework	Prediction
BTReport (gpt-oss:120B)	ROUGE: 0.37 BLEU: 0.24 RATESCR: 0.56 BERTSCR: 0.47 TB-F1: 0.33 **MASS EFFECT & VENTRICLES:** Approximately 12 mm of right-to-left midline shift is present, measured at the level of the falx cerebri above. The ventricles are asymmetrical, with compression of the right lateral ventricle secondary to tumor-related ependymal invasion; there is no overall ventricular enlargement. No tonsillar herniation is identified. **BRAIN/ENHANCEMENT:** A solitary right-sided lesion involving the temporal, cortical, and parietal lobes measures 7.1 x 5.6 x 5.3 cm (AP x TV x CC). The mass shows marked heterogeneous enhancement with a thick ($> 3mm$) enhancing margin; roughly 28% of the tumor volume enhances. A necrotic core comprises about 22% of the lesion volume. Deep white-matter invasion and cortical involvement are evident, and there is direct ependymal invasion of the adjacent lateral ventricle. The enhancing component does not cross the midline, but extensive vasogenic edema (84 mL, around 50% of total lesion volume) crosses the midline and surrounds the lesion, extending into the contralateral hemisphere. Multiple small enhancing satellite nodules are present along the cortical surface. The edema involves visual-association cortex, indicating eloquent brain (vision) involvement.

BTReport (LLaMa3:70B)

ROUGE: 0.42 BLEU: 0.29 RATESCR: 0.57 BERTSCR: 0.47 TB-F1: 0.34

MASS EFFECT & VENTRICLES: There is a maximum of 12 mm leftward midline shift at the level of the falx cerebri above. The ventricles are asymmetrical, with tumor-side ventricular asymmetry noted. There is no enlargement of the ventricles.

BRAIN/ENHANCEMENT: Within the right temporal, cortex, and parietal lobes, there is a solitary, predominantly enhancing mass measuring 7.1 x 5.6 x 5.3 cm. The lesion exhibits marked enhancement with a thick ($> 3mm$) enhancing margin. There are multiple small satellite lesions present. The enhancing region does not cross the midline. The lesion invades the cortex and deep white matter, with ependymal invasion also present. There is a large volume of surrounding vasogenic edema (83.6 mL), which crosses the midline. The proportion of edema is 49.74%. A central focus of necrosis is present, comprising approximately 21.99% of the lesion."

AutoRG-Brain

ROUGE: 0.32 BLEU: 0.25 RATESCR: 0.53 BERTSCR: 0.41 TB-F1: 0.25

After contrast administration, the lesion shows significant enhancement on T1-weighted images with unclear boundaries. The adjacent brain parenchyma shows significant edema in the brain pools and ventricles. After contrast administration, the lesion in the right temporal lobe demonstrates significant ring-like enhancement. The lesion measures approximately 82*111*82mm, with indistinct borders and a large amount of edema in the adjacent brain tissue, causing compression of the right lateral ventricle and deviation of midline structures to the left.

Seg-to-Exp

ROUGE: 0.21 BLEU: 0.16 RATESCR: 0.41 BERTSCR: 0.26 TB-F1: 0.0

The brain tumor segmentation results indicate that the tumor is primarily located in the right temporal and parietal regions, with a significant presence in the Temporal-to-Parietal (GapMap) region (40.03% of the tumor). The tumor also affects several surrounding regions, including Area STS2 (STS), Area STS1 (STS), Frontal-to-Temporal-II (GapMap), and Area TE 2.2 (STG).
Semantic segmentation of the tumor reveals three distinct categories: **Tumor Core** (red), **Peritumoral Edema** (yellow), and **GD-Enhancing Tumor** (green). The Tumor Core represents the most aggressive part of the tumor, while Peritumoral Edema indicates the area surrounding the tumor that is affected by edema. The GD-Enhancing Tumor region shows enhancement on Gadolinium (GD) MRI sequences, suggesting a highly vascularized and active part of the tumor.