ACQUISITION-AWARE PATCH AGGREGATION FOR 3D BRAIN MRI REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A brain MRI examination provides rich diagnostic information through multiple contrasts and series, but at the cost of substantial volumetric data. Leveraging such volumetric information for radiology report generation poses a challenge, as a vision–language model must accommodate the extensive volumetric information within a limited number of visual tokens. Existing approaches typically sample a subset of slices or compress visual representations before language generation. While necessary for tractability, these strategies may overlook where clinically relevant evidence lies: sampling can omit images containing important findings, while compression can obscure fine-grained local information. To address this, we introduce MR-Weave, a 3D brain MRI report generation model that weaves complementary evidence across acquisitions into a unified representation within a fixed visual token budget. MR-Weave groups features from all acquisitions by depth and performs *acquisition-aware patch aggregation*, adaptively weighting acquisition features at each patch position. We further introduce *axial geometry encoding*, which augments the fused tokens with physical position and scale information to support location- and size-aware descriptions. We evaluate MR-Weave on 2,000 studies from the co-registered release of MR-RATE using an LLM-as-a-judge protocol across five clinical dimensions: detection, faithfulness, disease labeling, localization, and severity. MR-Weave achieves the highest Overall score among the compared methods, outperforming the strongest baseline by 19%.

1 INTRODUCTION

Radiology report generation aims to produce textual descriptions of clinically relevant findings directly from medical images, with the potential to assist radiologists during report composition. Multi-contrast MRI makes this task particularly demanding. Each examination may contain multiple contrasts and repeated series, while the most informative contrast or series varies across findings and spatial location. Fig. 1 summarizes the complementary evidence provided by representative acquisitions (Bitar et al., 2006; Azhar & Chong, 2023). T2-weighted images can reveal fluid-related abnormalities, while fluid-attenuated inversion recovery (FLAIR) suppresses cerebrospinal fluid to better delineate periventricular abnormalities. Complementarity also exists across series of the same contrast: comparing pre- and post-contrast T1-weighted images helps distinguish true enhancement from intrinsically high signal (Ellingson et al., 2015). Consequently, no single acquisition is uniformly informative across an examination; its relevance depends on both the finding and spatial location.

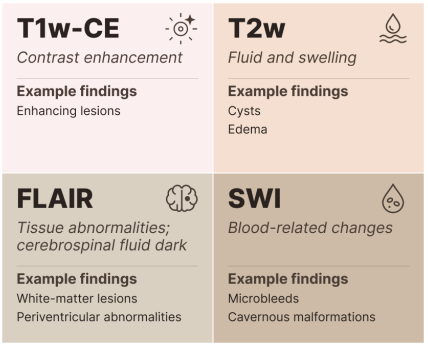
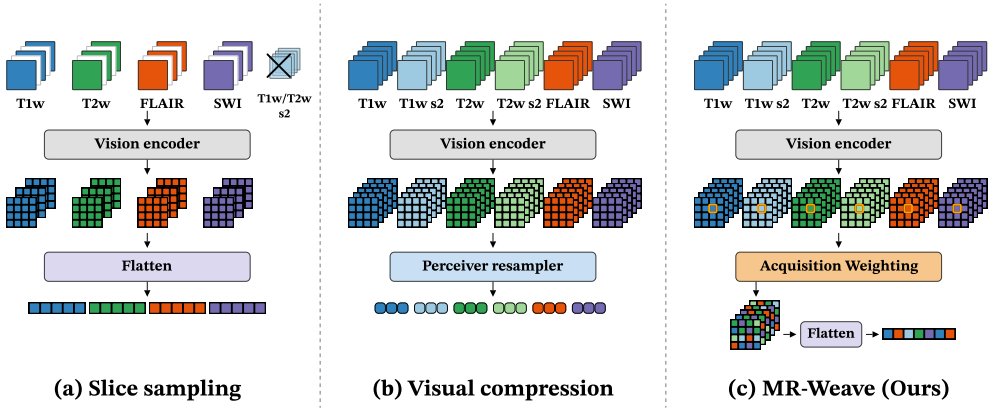


Figure 1: **Complementary MRI Acquisitions.** Different acquisitions highlight distinct findings, motivating acquisition-aware aggregation.

Leveraging such complementary information requires incorporating multiple acquisitions across the examination. In vision–language models, however, encoding a large number of slices produces correspondingly many visual tokens, making it impractical to provide all of them to the language

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



(a) Slice sampling **(b) Visual compression** **(c) MR-Weave (Ours)**

Figure 2: **Three strategies for representing an MRI study within a limited visual-token budget. (a) Slice sampling** (MedGemma 1.5) uniformly samples 85-slice across contrasts. A vision encoder and projector map each selected slice to a 16×16 token grid. These grids form the language model's visual input. **(b) Visual compression** (NeuroVFM) encodes each acquisition, then uses a Perceiver to compress its features into 64 latent tokens whose positions do not directly correspond to spatial locations in the input images. **(c) MR-Weave (ours)** groups features from all acquisitions by depth and adaptively weights acquisition features at each patch position, yielding one 16×16 token grid per depth bin. The amber squares indicate the same patch position across acquisitions.

model. Existing methods address this issue by sampling a subset of slices (MedGemma 1.5 (Sellergren et al., 2026)) or compressing visual features into latent tokens (NeuroVFM (Kondepudi et al., 2026)) before report generation. As illustrated in Fig. 2, these strategies reduce input length in different ways. Sampling discards information from unselected slices, whereas latent compression reduces the spatial detail retained from the input images.

While existing approaches reduce visual tokens through sampling or compression, we observe that information from multiple acquisitions need not be represented independently. MRI acquisitions can be spatially aligned to a common anatomical space, establishing correspondence across contrasts and series. This observation motivates combining features at corresponding spatial locations, with their contributions determined by local image content. Building on this idea, we propose MR-Weave, a model for 3D brain MRI report generation based on acquisition-aware patch aggregation. MR-Weave first groups slices from different acquisitions according to their axial positions. Within each resulting depth bin, it adaptively weaves features at corresponding patch locations across acquisitions into a single spatial representation. Each bin produces a single spatial token grid that integrates complementary evidence across acquisitions while maintaining the same maximum visual-token budget as the slice-sampling baseline. Axial geometry encoding further provides context for anatomical localization and size descriptions. We evaluate our approach on 2,000 MR-RATE test studies across five clinical dimensions, demonstrating consistent improvements over fine-tuned MedGemma 1.5.

Our contributions are threefold:

- We introduce MR-Weave, which uses *acquisition-aware patch aggregation* to weave features across acquisitions and *axial geometry encoding* to preserve spatial context, within a fixed visual-token budget.
- We formulate a five-axis evaluation protocol for 3D brain MRI report generation that accounts for finding coverage and distinguishes clinically relevant error types, with its sensitivity assessed through controlled report perturbations.
- We achieve an Overall score 19% higher than that of the strongest baseline on 2,000 MR-RATE test studies, with ablation studies and attention visualizations further demonstrating the contributions of individual components and acquisitions.

2 Related Work

Brain MRI report-generation systems have taken different approaches to grounding findings in clinically relevant evidence. AutoRG-Brain (Lei et al., 2026) uses anatomical and abnormality masks to connect generated descriptions to lesion topography; this explicit grounding improves interpretability, but its report module depends on region-level supervision that is not available in many report datasets. PIRTA (Lee et al., 2026) retrieves clinically similar DWI/ADC volumes and their paired reports to guide generation, reducing the burden on a direct image-to-text mapping but relying on the retrieval set and report similarity. Both approaches emphasize clinically meaningful evidence, yet neither addresses how to combine multiple co-registered contrasts and repeat series before generation.

Generalist medical vision–language models provide a complementary route to volumetric report generation. NeuroVFM (Kondepudi et al., 2026) learns neuroanatomical and pathological representations from large-scale clinical MRI and CT volumes and compresses each series through a volumetric encoder and Perceiver connector. However, this latent compression can weaken the correspondence between individual tokens and their anatomical locations. MedGemma 1.5 (Sellergren et al., 2026) is a general medical multimodal language-model backbone that can process medical images for image-to-text generation. It processes each study through a fixed sample of 2D slices, reducing the cost of passing volumetric data to a language model but potentially leaving informative slices unavailable. Unlike slice subsampling or aggressive latent compression, which can discard diagnostically relevant information, our approach selectively aggregates patch-level evidence while keeping the number of input tokens equal to that of the slice-sampling baseline.

Evaluation methods for report generation have likewise moved from surface similarity toward clinical content. Lexical and embedding-based metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee & Lavie, 2005), and BERTScore (Zhang et al., 2020), remain common, but they can reward wording overlap while missing clinically important omissions, false statements, or localization errors (Yu et al., 2023). Classifier-based measures, including CheXbert F1 (Smit et al., 2020) and RadBERT-based classifiers (Yan et al., 2022; Hamamci et al., 2024), are constrained by predefined label spaces; entity- and graph-based measures such as RadGraph (Jain et al., 2021), RaTEScore (Zhao et al., 2024), and RadGraph-XL (Delbrouck et al., 2024) capture structured relations but depend on extraction schemas and matching accuracy. Brain-specific metrics add anatomical and pathology labels, yet remain tied to fixed keyword or label sets (Li et al., 2025a; Shi et al., 2023; Barone et al., 2026). LLM-based evaluators such as GREEN (Ostmeier et al., 2024), CheXprompt (Zambrano Chaves et al., 2025), FineRadScore (Huang et al., 2024), and RadReason (Li et al., 2025b) assess clinical meaning more directly through error categories or severity judgments. Following this trend, we evaluate generated reports along five clinical dimensions and test the evaluator's sensitivity to controlled clinical errors.

3 Method

3.1 Problem Formulation and Overview

Given a brain MRI study comprising multiple volumetric acquisitions from a single examination, our goal is to generate the corresponding radiology report. The study may include T1-weighted (T1w), T2-weighted (T2w), fluid-attenuated inversion recovery (FLAIR), and susceptibility-weighted imaging (SWI). We refer to each imaging type as a *contrast*, and each three-dimensional volume as an *acquisition*. Multiple acquisitions of the same contrast may be present in a single examination, for example, T1w scans acquired before and after contrast-agent administration; we refer to each such acquisition as a *series* of that contrast. We index an acquisition by its contrast $m \in \mathcal{M} = \{\text{T1w}, \text{T2w}, \text{FLAIR}, \text{SWI}\}$ and series index $s \in \{1, \dots, N_m\}$, where N_m is the number of series of contrast m present in the study. A study $\mathcal{X}$ is then written as

$$\mathcal{X} = \left\{ \mathbf{V}^{(m,s)} \in \mathbb{R}^{S \times H \times W} \;:\; m \in \mathcal{M}_{\mathcal{X}},\; s \in \{1, \dots, N_m\} \right\}, \tag{1}$$

where all volumes are co-registered to a shared voxel grid. Here, S denotes the number of slices along the superior–inferior axis, and $H \times W$ the shared in-plane grid, such that voxels with the same index correspond spatially across acquisitions. A study therefore contains $N = \sum_m N_m$ acquisitions and NS slices in total. Both the contrast set $\mathcal{M}_{\mathcal{X}} \subseteq \mathcal{M}$ and the number of series N_m for each contrast vary across studies.

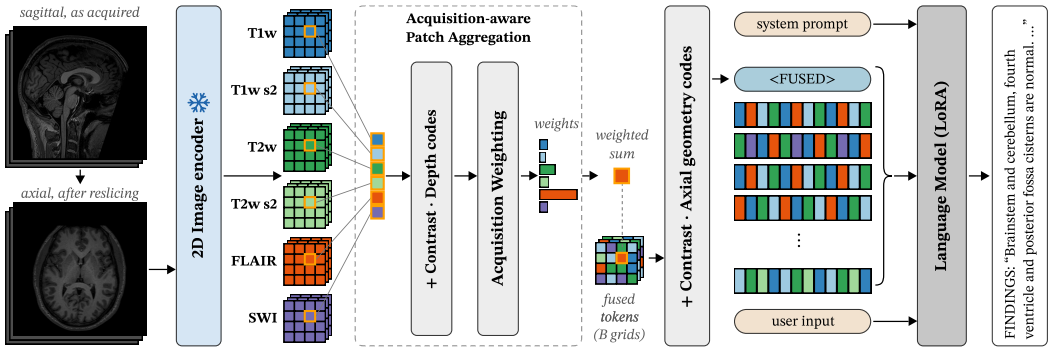


Figure 3: **Overview of the method.** Axial reslicing places studies in a common anatomical orientation, and the frozen image encoder encodes slices from each acquisition. *Acquisition-aware patch aggregation* augments candidate features with contrast and normalized-depth codes, then assigns acquisition weights at each patch position (amber) to produce a fused token. Repeating this operation across patch positions and axial locations yields B fused token grids. *Axial geometry encoding* augments the fused tokens with in-plane position and axial geometry. The LoRA-adapted language model receives the resulting grids together with the system prompt, `<FUSED>` marker, and user instruction to generate the report.

Representing all NS slices independently would produce a large number of visual tokens, whereas restricting the input to a subset of slices may leave information from other acquisitions unavailable to the language model. Our goal is therefore to incorporate information across acquisitions within a fixed visual token budget while retaining the spatial correspondence between visual tokens and anatomy. Fig. 3 provides an overview of our approach. We first reslice the volumes into a common axial orientation and encode each slice into visual tokens (§3.2). Acquisition-aware patch aggregation then combines features across contrasts and series at corresponding spatial locations (§3.3). Finally, we apply axial geometry encoding and feed the resulting tokens into the language model to generate the report (§3.4).

3.2 Visual Token Preparation

MRI acquisitions are acquired in different planes, including axial, sagittal, and coronal orientations. Without reorientation, the same grid axis would represent different anatomical directions across studies; for sagittal slices, laterality would lie along the slice axis rather than within the image. We therefore reformat all co-registered volumes into axial slices before encoding (Fig. 3), giving all studies a consistent orientation and an inferior-to-superior slice order.

With this common orientation, we partition the axial extent into $B = \min(S, 85)$ equal-height bins shared across all acquisitions, where S is the number of axial slices in the shared voxel grid after reslicing. For each bin b, candidate slices from multiple acquisitions form the set $\mathcal{C}_b$ for subsequent patch aggregation. Details of candidate construction are provided in Appendix B.2.

We then encode each candidate frame $f \in \mathcal{C}_b$ using MedGemma 1.5's frozen vision encoder and projector. The same encoder is applied to every frame of every bin, independently of its contrast m_f and series s_f. For each frame, it produces a token grid $\mathbf{t}_f \in \mathbb{R}^{T \times D}$, where $T = 256$ tokens form a 16×16 grid and $D = 2560$. The *acquisition-aware patch aggregation* stage then combines the candidate grids within each bin into a single grid, so that a study of any size is represented by $B \times T$ tokens of dimension D regardless of how many acquisitions it contains.

3.3 Acquisition-Aware Patch Aggregation

Acquisition weighting. At a patch position $p \in \{1, \ldots, T\}$, we estimate how much each candidate frame $f \in \mathcal{C}_b$ contributes to the fused feature. Let $\mathbf{t}_{f,p} \in \mathbb{R}^D$ denote the visual feature at position p in frame f. Before computing the weight, we augment the feature with the frame's contrast and depth information:

$$\mathbf{x}_{f,p} = \mathbf{t}_{f,p} + \mathbf{e}^{\text{gate}}_{\text{contrast}}(m_f) + \mathbf{e}_{\text{depth}}(\boldsymbol{\delta}_f), \qquad \boldsymbol{\delta}_f = \left(\Delta z_f^{\text{bin}},\ \Delta z_f^{\text{study}}\right), \tag{2}$$

where m_f identifies the frame's contrast and $\mathbf{e}^{\text{gate}}_{\text{contrast}}(m_f) \in \mathbb{R}^D$ is retrieved from a learned embedding table. The two entries of $\boldsymbol{\delta}_f$ represent the frame's normalized depth within its bin and within the study's axial extent, respectively. Unlike the discrete contrast embedding, the depth embedding is obtained from the continuous coordinates through a learned affine projection, $\mathbf{e}_{\text{depth}}(\boldsymbol{\delta}_f) = \mathbf{W}_z\boldsymbol{\delta}_f + \mathbf{b}_z$, where $\mathbf{W}_z \in \mathbb{R}^{D\times 2}$ and $\mathbf{b}_z \in \mathbb{R}^D$. Both embeddings have the same dimension as the visual feature and are shared across all patch positions within a frame. Exact depth normalization is specified in Appendix B.3.

The augmented feature $\mathbf{x}_{f,p}$ is then passed through layer normalization (LN) and a two-layer MLP to obtain a scalar score $\ell_{f,p}$. The same scoring network is shared across all frames and patch positions. A softmax over candidates within the bin then converts the scores into weights:

$$\ell_{f,p} = \text{MLP}\big(\text{LN}(\mathbf{x}_{f,p})\big) \qquad w_{f,p} = \frac{\exp \ell_{f,p}}{\sum_{f'\in\mathcal{C}_b} \exp \ell_{f',p}}. \tag{3}$$

For each position p, these weights are nonnegative and sum to one over $f \in \mathcal{C}_b$. They therefore specify how strongly each candidate contributes at that position, allowing the aggregation to draw differently from contrasts and series across spatial locations.

Feature fusion. We use the weights $w_{f,p}$ to combine candidate features at the same patch position. To retain their contrast information, we include a learned contrast embedding in the weighted sum, producing a fused token $\mathbf{u}_{b,p} \in \mathbb{R}^D$:

$$\mathbf{u}_{b,p} = \sum_{f\in\mathcal{C}_b} w_{f,p}\,\big(\mathbf{t}_{f,p} + \mathbf{e}^{\text{fuse}}_{\text{contrast}}(m_f)\big). \tag{4}$$

Here, $\mathbf{e}^{\text{fuse}}_{\text{contrast}}(m_f) \in \mathbb{R}^D$ is a learned embedding for the contrast of frame f. The resulting token combines visual features across candidates while retaining information about their contrast identities. For example, when a patch assigns greater weight to a FLAIR candidate, both the visual feature and contrast embedding of the FLAIR candidate contribute proportionally more to the fused token. Physical position and scale information is subsequently added to form the language-model input (§3.4).

Repeating this operation over all T patch positions produces one fused grid per bin, regardless of the number of candidate frames. Because the weights sum to one, each fused token is a weighted average of the candidate features with their contrast embeddings added. Most patches retain a mixture of candidates rather than selecting a single frame.

3.4 Axial Geometry Encoding

To help the model interpret anatomical laterality and lesion size, we augment the fused tokens with physical position and scale information on the common axial grid; we refer to this step as axial geometry encoding.

Token-level in-plane position. We assign each token a physical location within the axial image, expressed as an offset from the image center. Each study has a square in-plane extent with a study-specific side length L in millimeters, shared across all slices. The spacing between adjacent positions on the 16×16 token grid is therefore $\Delta_{xy} = L/16$. For a token p at column i and row j, where $i, j \in \{0, \dots, 15\}$, its center coordinates are $x_p = (i-7.5)\Delta_{xy}$ and $y_p = (j-7.5)\Delta_{xy}$. Subtracting 7.5 places the origin at the center of the token grid.

To represent these locations at multiple spatial scales, we augment the two coordinates with sine and cosine features at eight frequencies, $\omega_k = 2^k\pi/100$ for $k = 0, \dots, 7$. The resulting vector $\boldsymbol{\phi}(p) \in \mathbb{R}^{34}$ is mapped to a D-dimensional position embedding $\mathbf{e}_{xy}(p)$ by a learned affine projection:

$$\begin{aligned}\boldsymbol{\phi}(p) &= \big[x_p, y_p, (\sin\omega_k x_p)_{k=0}^7, (\sin\omega_k y_p)_{k=0}^7, (\cos\omega_k x_p)_{k=0}^7, (\cos\omega_k y_p)_{k=0}^7\big] \in \mathbb{R}^{34},\\ \mathbf{e}_{xy}(p) &= \mathbf{W}_{xy}\,\boldsymbol{\phi}(p) + \mathbf{b}_{xy}.\end{aligned} \tag{5}$$

Lower frequencies capture gradual spatial variation, while higher frequencies encode finer positional differences. Together with the original millimeter coordinates, these features provide a multi-scale representation of each patch's physical in-plane location.

Bin-level position and scale. The in-plane code locates each token within an axial slice. We complement it with a bin-level code that captures axial position and physical scale. For bin b, we

define $\mathbf{g}_b = [z_b, \Delta z_b, \sigma, \Delta_{xy}]$. Here, z_b is the bin's offset from the midpoint of the volume along the z-axis, and Δz_b is the axial extent spanned by the candidate slices in the bin. The interval σ is the median distance between consecutive candidate slice locations across the study's axial extent, while Δ_{xy} is the physical distance between adjacent token-grid positions within a slice. All four quantities are measured in millimeters. We normalize them using fixed scaling constants $\mathbf{c}$ and project the resulting vector to a D-dimensional embedding shared across all tokens in the bin:

$$\widetilde{\mathbf{g}}_b = \mathbf{g}_b \oslash \mathbf{c}, \qquad \mathbf{e}_{\text{geom}}(b) = \mathbf{W}_g \widetilde{\mathbf{g}}_b + \mathbf{b}_g, \tag{6}$$

where $\oslash$ denotes element-wise division. The fixed scaling brings the four features to comparable numerical ranges while preserving their physical meaning.

Language-model input. We add the in-plane position and bin-level geometry embeddings to obtain the final visual token at patch p in bin b:

$$\mathbf{h}_{b,p} = \mathbf{u}_{b,p} + \mathbf{e}_{xy}(p) + \mathbf{e}_{\text{geom}}(b). \tag{7}$$

The input sequence consists of the system prompt, a <FUSED> marker, the $B \times T$ visual tokens, and a one-line report-generation instruction. The full template is provided in Appendix C.1. The LoRA-adapted language model then generates the report autoregressively.

Training. The vision encoder and projector remain frozen. We jointly train the weighting MLP, contrast embedding tables, depth, position, and geometry projections, and the language model's LoRA adapters using the report-generation loss. Training settings are given in Appendix B.4.

4 Experiments

Our experiments address three research questions. **RQ1.** Does MR-Weave improve 3D brain MRI report generation compared with existing approaches? (§4.2) **RQ2.** To what extent do acquisition-aware patch aggregation and axial geometry encoding improve report quality? (§4.3) **RQ3.** Does the model attend to relevant MRI acquisitions and pathology regions when generating individual findings? (§4.4)

4.1 Experimental Setup

Datasets. We use MR-RATE (Forithmus, 2026), an openly released corpus (CC BY-NC-SA) of 98,334 brain and spine MRI studies from 83,425 patients with paired radiology reports, partitioned into patient-disjoint training, validation, and test splits. We use the co-registered release, in which every series of a study is resampled onto the grid of its T1w reference volume so that all contrasts share one coordinate frame. The model takes T1w, T2w, FLAIR, and SWI acquisitions as input and generates the report. We exclude reports that mention spine anatomy and remove references to prior examinations, which cannot be verified from the current images, leaving 81,607 training studies and a non-overlapping test set of 2,000 (Appendix B.1).

Baselines. We select recent methods that generate brain MRI reports as our baselines. MedGemma 1.5 shares our pretrained backbone and provides a slice-sampling reference. NeuroVFM represents volumetric encoding with latent compression, while AutoRG-Brain provides a brain-specific approach grounded in anatomical and anomaly segmentation. We compare MedGemma 1.5 and NeuroVFM in both zero-shot and fine-tuned settings, while evaluating AutoRG-Brain only in the zero-shot setting because MR-RATE does not provide the pathology region annotations required to fine-tune its report-generation module. Baseline implementation details are provided in Appendix B.5.

Training and inference. We use MedGemma 1.5 (4B) as the backbone and jointly train the fusion modules and rank-16 language-model LoRA adapters, keeping the vision encoder and projector frozen. Training runs for two epochs on four NVIDIA B200 GPUs with a global batch size of 32 studies (eight per GPU), using AdamW and a warmup–stable–decay (WSD) learning-rate scheduler. At test time, the final checkpoint generates reports for the 2,000 held-out studies using greedy decoding. Further settings are provided in Appendix B.4.

Evaluation. Following reference-based LLM evaluators such as GREEN and FineRadScore (Ostmeier et al., 2024; Huang et al., 2024), we use Gemini 2.5 Flash as a judge to compare generated and

Table 1: LLM-based evaluation on the 2,000 test studies. Det.: pathology detection; Faith.: faithfulness; DisLab.: disease labeling; Loc.: localization; Sev.: severity. Overall is the mean of the five axis columns; higher is better. **Bold** and underlined denote the best and second-best results, respectively.

Method	Det.	Faith.	DisLab.	Loc.	Sev.	**Overall**
AutoRG-Brain (zero-shot)	0.961	0.804	0.256	0.593	0.479	0.619
MedGemma 1.5 (zero-shot)	0.732	2.062	0.569	0.458	0.438	0.852
MedGemma 1.5 (fine-tuned)	<u>1.569</u>	<u>2.825</u>	<u>1.327</u>	<u>1.118</u>	<u>0.767</u>	<u>1.521</u>
NeuroVFM (zero-shot)	0.853	**2.949**	0.652	0.473	0.392	1.064
NeuroVFM (fine-tuned)	0.654	2.637	0.566	0.451	0.292	0.920
MR-Weave (Ours)	**1.981**	2.739	**1.737**	**1.375**	**1.223**	**1.811**

ground-truth reports from the same study. We evaluate the generated reports along five clinical dimensions: pathology detection, faithfulness, disease labeling, localization, and severity. The judge's temperature is set to 0 to reduce sampling variability. Each dimension is graded on a 0–5 scale in half-point steps: *pathology detection* assesses whether ground-truth abnormalities are recognized in the generated report; *faithfulness* penalizes fabricated findings and explicit claims of normality that contradict abnormalities in the ground-truth findings; and *disease labeling*, *localization*, and *severity* assess the correctness of each reported finding's diagnosis, anatomical location and laterality, and severity (including size and extent), respectively. Overall is the equally weighted mean of the five axis-level scores. Details of the evaluation rubric and score aggregation are provided in Appendix B.7.

4.2 Comparative Results

We compare the clinical content of generated reports across MedGemma 1.5, NeuroVFM, AutoRG-Brain, and our method on the same 2,000 test studies using the five-axis evaluation rubric (Table 1).

Our model achieves the highest overall score among the compared methods, with the best scores in pathology detection, disease labeling, localization, and severity. Compared with fine-tuned MedGemma 1.5, the strongest baseline overall, our method improves Overall from 1.521 to 1.811 (+0.290; a 19.1% relative increase). It also achieves the highest scores across all standard report-generation metrics (Appendix A.1), showing consistent improvements in both clinical-content and text-similarity evaluations.

Fine-tuning improves MedGemma 1.5 across all five axes, but its Overall score remains below ours (1.521 versus 1.811). For NeuroVFM, fine-tuning lowers scores across all five axes, with Overall falling from 1.064 to 0.920. The contrast between the LLM-as-a-judge results in Table 1 and the conventional text-similarity results in Table 3 is particularly evident for NeuroVFM. Fine-tuning substantially improves its conventional text-similarity scores (Table 3), while its pathology-detection score in the LLM-as-a-judge evaluation (Table 1) decreases from 0.853 to 0.654 and the other clinical axes also decline. Standard report-generation metrics reward overlap in wording or extracted clinical entities and relations, so matching templated normal findings can raise scores even when abnormalities are missed. The divergent results suggest that fine-tuning makes NeuroVFM's outputs more similar to the reference reports in style and recurring textual structure without improving recognition of pathological findings. Zero-shot NeuroVFM scores higher than our model on faithfulness (2.949 versus 2.739), but detects fewer reference abnormalities (0.853 versus 1.981). This trade-off is consistent with conservative reporting: omissions alone are not penalized on the faithfulness axis, so a high score does not necessarily imply complete reporting. Considering all five dimensions together, our model achieves the highest Overall score, balancing finding coverage with the accuracy of the reported findings. Appendix A.2 provides controlled-perturbation experiments that validate the evaluator's sensitivity to clinically relevant errors.

Table 2: **Ablation study of fusion strategy and spatial encoding.** We assess the effects of alternative fusion strategies, spatial encodings, and their combination. Slice sampling corresponds to fine-tuned MedGemma 1.5 in Table 1. The highlighted row is our final model, MR-Weave. **Bold** and underlined denote the best and second-best results, respectively.

Fusion Strategy	Axial Reslicing	Spatial Encoding	Det.	Faith.	DisLab.	Loc.	Sev.	**Overall**
Slice Sampling	–	–	1.569	2.825	1.327	1.118	0.767	1.521
Latent-channel concat.	–	–	1.093	2.669	0.950	0.712	0.806	1.246
Image-channel concat.	–	–	1.650	**2.867**	1.453	1.106	1.035	1.622
Patch-wise hard selection	–	–	1.753	2.858	1.465	1.231	0.930	1.647
Patch-wise mean	–	–	1.794	2.595	1.582	1.214	1.124	1.662
Patch-wise softmax	–	–	1.901	2.624	1.665	1.277	1.165	1.726
Patch-wise softmax	–	RAS	1.910	2.557	1.668	1.299	1.204	1.727
Patch-wise softmax	✓	Axial	**1.981**	2.739	**1.737**	**1.375**	**1.223**	**1.811**

4.3 Ablation Studies

We conduct ablation studies to assess how much acquisition-aware patch aggregation and axial geometry encoding contribute to report quality (Table 2). Implementation details are provided in Appendix B.6.

Fusion strategies. To compare patch aggregation with alternative fusion strategies, we consider slice sampling, image-channel concatenation, latent-channel concatenation, patch-wise hard selection, and patch-wise mean aggregation. We refer to our learned softmax weighting as *patch-wise softmax* to distinguish it from the two alternative patch-wise strategies. Slice sampling is the fine-tuned MedGemma 1.5 baseline in Table 1; it distributes an 85-slice budget across available contrasts and passes the sampled slices to the vision encoder. Image-channel concatenation concatenates contrast images along the channel dimension to form the input to the vision encoder, whereas latent-channel concatenation concatenates projected contrast features along the channel dimension at corresponding patch positions, followed by a learned fusion projection. Patch-wise hard selection and mean aggregation represent two limiting cases of patch aggregation: hard selection retains only the highest-weighted candidate at each patch position, whereas mean aggregation assigns equal weights to all candidates.

Even without reslicing or spatial encoding, patch aggregation achieves the highest Overall score of 1.726 among the six fusion strategies and scores highest in detection, disease labeling, localization, and severity. Among the three patch-wise strategies, patch-wise softmax outperforms both hard selection (1.647) and mean aggregation (1.662) in Overall, supporting adaptive soft weighting over either single-candidate selection or uniform averaging.

Spatial encoding. To assess how spatial information benefits patch aggregation, we compare RAS positional encoding on the original slice planes with axial geometry encoding. RAS (right–anterior–superior) encoding represents token positions along the anatomical right–left, anterior–posterior, and superior–inferior axes, whereas the resliced configuration encodes physical position and scale on a common axial grid. RAS encoding alone yields little improvement in Overall, as modest gains in detection, disease labeling, localization, and severity are offset by lower faithfulness. Adding axial geometry encoding (*Axial* in Table 2) achieves the highest Overall score of 1.811, outperforming patch aggregation both without spatial encoding and with RAS encoding. This configuration is our final model, MR-Weave. It also improves all five evaluation axes over both configurations, supporting the combined use of reslicing and physical geometry codes.

4.4 Attention-Based Evidence Analysis

To assess whether our model attends to relevant MRI acquisitions and pathology regions when generating individual findings, we analyze its attention during report generation in three steps. First, we compute attention to each fused patch token during generation of a finding sentence, averaging over attention heads and global-attention layers and summing over the sentence's generated tokens. This attention identifies spatial locations without distinguishing the acquisitions that contributed

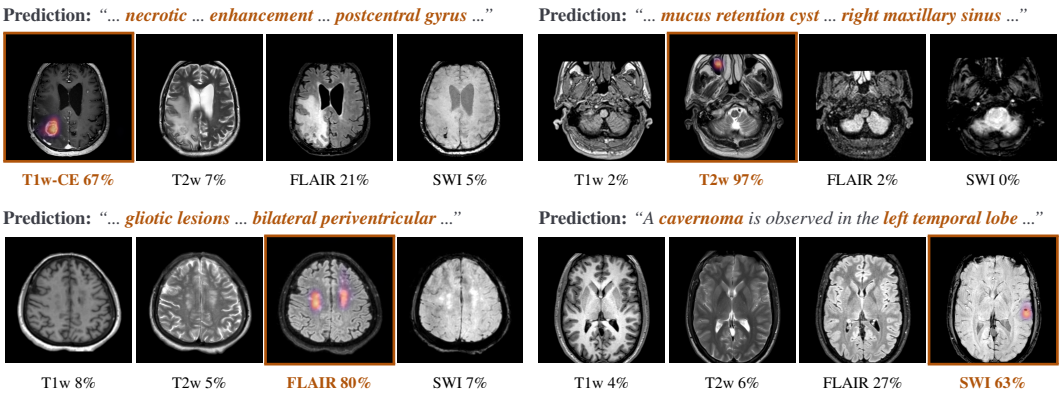


Figure 4: **Attention-Based Evidence Analysis.** This figure visualizes the MRI evidence to which the model attends while generating findings, using four examples with co-registered contrasts. Across the examples, the model attends to the contrast most suited to the described finding and concentrates on the reported lesion location, indicating that the fused representation preserves lesion-relevant visual evidence. Percentages indicate each contrast's share of the visual evidence, summed over all series of that contrast; the slice shown is the series that carried most of that share, which for the first example is the post-contrast T1w acquisition (T1w-CE). Image brightness is adjusted for visualization.

to each fused token. Second, within the depth bin receiving the most attention, we multiply each patch's attention by the patch aggregation weights of the acquisitions contributing to it. Third, we sum the weighted attention over patches for each contrast and normalize across contrasts to obtain percentages summing to 100%, which are overlayed to the corresponding images in Figure 4.

Figure 4 presents four qualitative examples in which attention concentrates on acquisitions suited to characterizing the described lesions: contrast-enhanced T1w for a necrotic, enhancing mass (67%), T2w for a maxillary sinus mucus retention cyst (97%), FLAIR for periventricular gliotic lesions (80%), and SWI for a cavernoma (63%). The highlighted regions also correspond to the anatomical locations described in the generated report. Together, these examples show that the model attends to acquisitions and spatial regions relevant to each finding. Beyond the report itself, this gives the model a degree of interpretability: because the aggregation weights and the attention are indexed by acquisition and patch position, each generated sentence can be traced back to the acquisitions and image regions that supported it.

5 Conclusion

We presented MR-Weave, an acquisition-aware patch aggregation model for 3D brain MRI report generation. By combining co-registered contrast and series features at corresponding patch locations, patch aggregation integrates complementary acquisitions within a fixed visual-token budget. Axial geometry encoding further provides physical spatial information for report generation. On 2,000 MR-RATE test studies, our method outperforms the strongest current baselines, achieving the highest Overall score. Qualitative attention visualizations further show that the model attends to finding-relevant contrasts and anatomical regions during generation, providing interpretability beyond report text alone. Together, these results suggest a direction for 3D MRI report generation: rather than treating volumetric input reduction as a global sampling or compression problem, models can preserve local evidence by selectively integrating spatially aligned information across acquisitions. Future work will extend this direction with independent evaluation by radiologists.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code development, debugging, and manuscript editing. As part of the experimental pipeline, Gemini 2.5 Flash was used to rewrite prior-examination references in reports, and evaluate generated reports with the stated rubric. These uses are described in §4.1 and Appendices B.7 and C. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The method is described in §3 and the experimental setup in §4.1, with evaluation and statistical procedures in Appendix B.7, data preparation, implementation, and training details in Appendix B, and full prompts in Appendix C. Code will be released upon publication.

REFERENCES

Syifa Azhar and Le Roy Chong. Clinician's guide to the basic principles of MRI. *Postgraduate Medical Journal*, 99(1174):894–903, 2023. doi: 10.1136/pmj-2022-141998. (Cited on page 1.)

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72. Association for Computational Linguistics, 2005. URL https://aclanthology.org/W05-0909/. (Cited on page 3.)

Mariano Barone, Francesco Di Serio, Giuseppe Riccio, Antonio Romano, Marco Postiglione, Antonino Ferraro, and Vincenzo Moscato. Brain3D: Brain report automation via inflated vision transformers in 3D. *arXiv preprint arXiv:2602.22098*, 2026. URL https://arxiv.org/abs/2602.22098. (Cited on page 3.)

Richard Bitar, General Leung, Richard Perng, Sameh Tadros, Alan R. Moody, Josee Sarrazin, Caitlin McGregor, Monique Christakis, Sean Symons, Andrew Nelson, and Timothy P. Roberts. MR pulse sequences: What every radiologist wants to know but is afraid to ask. *RadioGraphics*, 26 (2):513–537, 2006. doi: 10.1148/rg.262055063. (Cited on page 1.)

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12902–12915. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.765. URL https://aclanthology.org/2024.findings-acl.765/. (Cited on page 3.)

Benjamin M. Ellingson, Martin Bendszus, Jerrold Boxerman, et al. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro-Oncology*, 17(9):1188–1198, 2015. doi: 10.1093/neuonc/nov095. (Cited on page 1.)

Forithmus. MR-RATE: A vision-language foundation model and dataset for magnetic resonance imaging. https://huggingface.co/datasets/Forithmus/MR-RATE, 2026. Dataset release; paper and author list not yet published as of September 2026. TODO update when the paper appears. (Cited on page 6.)

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated radiology report generation for 3D medical imaging. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15012 of *Lecture Notes in Computer Science*, pp. 476–486. Springer, 2024. doi: 10.1007/978-3-031-72390-2_45. (Cited on page 3.)

Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. FineRad-Score: A radiology report line-by-line evaluation technique generating corrections with severity scores. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of

Proceedings of Machine Learning Research. PMLR, 2024. URL https://proceedings.mlr.press/v252/huang24a.html. (Cited on pages 3 and 6.)

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2106.14463. (Cited on page 3.)

Akhil Kondepudi, Akshay Rao, Chenhui Zhao, Yiwei Lyu, Samir Harake, Soumyanil Banerjee, Jacob Ogle, Rushikesh Joshi, Anna-Katharina Meissner, Xinhai Hou, Cheng Jiang, Asadur Chowdury, Ashok Srinivasan, Brian Athey, Vikas Gulani, Aditya Pandey, Honglak Lee, and Todd Hollon. Health system learning enables generalist neuroimaging models. *Nature Medicine*, 32 (8):2831–2837, 2026. doi: 10.1038/s41591-026-04497-1. Code: https://github.com/MLNeurosurg/neurovfm. (Cited on pages 2 and 3.)

Junhyeok Lee, Yujin Oh, Dahyoun Lee, Hyon Keun Joh, Minchul Kim, Chul-Ho Sohn, Sung Hyun Baik, Cheol Kyu Jung, Jung Hyun Park, Kyu Sung Choi, Byung-Hoon Kim, and Jong Chul Ye. Improving factuality of 3D brain MRI report generation with paired image-domain retrieval and text-domain augmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, 2026. URL https://arxiv.org/abs/2411.15490. Accepted at MICCAI 2026; final proceedings metadata not yet available as of September 2026. (Cited on page 3.)

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. *IEEE Journal of Biomedical and Health Informatics*, 30(7), 2026. (Cited on page 3.)

Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, Shih-Jen Chen, Jiing-Feng Lirng, Kai-Wei Chang, and Shih-Hwa Chiou. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. *Nature Communications*, 16(1):2258, 2025a. doi: 10.1038/s41467-025-57426-0. (Cited on page 3.)

Yingshu Li, Yunyi Liu, Lingqiao Liu, Lei Wang, and Luping Zhou. RadReason: Radiology report evaluation metric with reasons and sub-scores. *arXiv preprint arXiv:2508.15464*, 2025b. URL https://arxiv.org/abs/2508.15464. Preprint; no peer-reviewed publication identified as of September 2026. (Cited on page 3.)

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/. (Cited on page 3.)

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/. (Cited on pages 3 and 6.)

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*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/. (Cited on page 3.)

Andrew Sellergren et al. MedGemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026. (Cited on pages 2 and 3.)

Yanzhao Shi, Junzhong Ji, Xiaodan Zhang, Liangqiong Qu, and Ying Liu. Granularity matters: Pathological graph-driven cross-modal alignment for brain CT report generation. In *Proceedings*

of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 6617–6630, Singapore, 2023. Association for Computational Linguistics. (Cited on page 3.)

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117/. (Cited on page 3.)

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022. doi: 10.1148/ryai.210258. (Cited on page 3.)

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023. doi: 10.1016/j.patter.2023.100802. (Cited on page 3.)

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025. doi: 10.1038/s41467-025-58344-x. (Cited on page 3.)

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. (Cited on page 3.)

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*, pp. 15004–15019. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.836. URL https://aclanthology.org/2024.emnlp-main.836/. (Cited on page 3.)