Enhancing Brain MRI Anomaly Detection and Reasoning with ROI Rethink and Synthetic Data

Shangkun Li[1*], Jie Xu[1*], Yi Guo[1], Zeju Li[1†], and Yuanyuan Wang[1†]

College of Biomedical Engineering, Fudan University, Shanghai, China
{22307130255, xujie23}@m.fudan.edu.cn, {guoyi, zejuli, yywang}@fudan.edu.cn

Abstract. Medical vision-language models typically generate diagnoses through single-pass inference without indicating which image regions support their conclusions. This lack of spatial grounding limits clinical utility: outputs cannot be audited, and models may hallucinate findings on normal scans. We present **BrReMark** (**Br**ain **Re**think via ROI **Mark**ing), a framework that introduces explicit region marking into brain MRI diagnosis. The model first generates hypotheses about potential abnormalities and grounds them through explicit bounding box marking, then verifies conclusions by re-examining the marked evidence. Training combines supervised fine-tuning on structured reasoning trajectories with reinforcement learning using a composite reward over localization accuracy and diagnostic reasoning. Furthermore, we integrate a domain randomization-based pathology synthesis augmentation strategy to improve the model's generalizability to out-of-distribution (OOD) data. On internal benchmark, BrReMark improves mAP_{50} from 0.74% to 37.54% compared to the base model, while achieving 21.57 Clinical F1 and 45.26% diagnostic accuracy. On NOVA OOD benchmark, it also achieves competitive overall performance with a 45.7% reduction in false positives compared to the state-of-the-art, indicating reduced hallucination on rare pathologies. These findings suggest that explicit hypothesis-verification grounding is a practical path toward trustworthy open-ended brain MRI diagnosis across both in-distribution and OOD settings.

Keywords: Brain MRI · Medical VLMs · Reasoning

1 Introduction

Trustworthy AI-assisted diagnosis requires not only accurate predictions but also auditable reasoning that clinicians can verify [18]. Medical vision-language models (VLMs) (e.g., LLaVa-Med [15], HuatuoGPT-Vision [7], and Lingshu [24]) have achieved strong performance on radiology dialogue and visual question answering, yet their clinical utility remains limited by a fundamental trustworthiness gap: models generate conclusions without indicating which image regions

[*] Equal contribution. [†] Corresponding authors.
Code is available at https://github.com/fdu-farm/BrReMark.

support the diagnosis, making outputs impossible to audit and prone to hallucination on normal scans [25]. This issue is particularly acute for brain MRI, where lesion location directly informs differential diagnosis and the prevalence of rare neurological diseases introduces severe out-of-distribution challenges [5].

To bridge this trustworthiness gap, the emerging "Thinking with Images" paradigm [9,22,27] offers a promising solution by shifting from passive observation to active perception. Unlike traditional models that rely on single-pass inference, this paradigm facilitates iterative interaction with visual data. This multi-turn reasoning process allows the model to formulate intermediate hypotheses, zoom in on diagnostically salient regions, and perform self-correction. Consequently, the final diagnostic decision is supported by a verifiable chain of visual evidence, aligning the model's behavior with clinical standards. Unlike fTSPL, which generates text from fMRI activations to assist representation learning and prediction [23], BrReMark explicitly grounds lesion locations in structural MRI and produces auditable ROI markings to support diagnosis.

Despite recent advancements, several critical limitations hinder its clinical application in brain MRI diagnosis. Primarily, prevailing benchmarks like VQA-RAD [13] and SLAKE [17] emphasize closed-ended accuracy, which inherently misaligns with the open-set nature of real-world clinical scenarios. Consequently, open-ended anomaly detection remains largely underexplored. Furthermore, constrained by these closed-set paradigms and limited data, existing models tend to overfit to known distributions, severely compromising their ability to generalize to out-of-distribution (OOD) data [5]. Developing novel methods capable of handling open-set clinical complexities is essential for real-world VLM deployment.

To this end, we present **BrReMark** (**B**rain **Re**think via ROI **Mark**ing), a novel "Think with Images" framework that tackles open-set OOD clinical challenges by explicitly mirroring the human diagnostic workflow. Specifically, BrReMark employs a progressive two-stage training paradigm: Supervised Fine-Tuning (SFT) followed by Group Relative Policy Optimization (GRPO [20])-based Reinforcement Learning (RL). Our main contributions are fourfold:

1. **Interactive "Mark-and-Rethink" Trajectory via SFT**: Rather than relying on standard QA pairs, this design elicits an auditable reasoning process: BrReMark first hypothesizes and explicitly marks suspicious regions of interest (ROIs), and subsequently uses these visual cues to "rethink" and verify initial findings. This explicit trajectory empowers it to navigate complex open-set clinical scenarios inherently neglected by closed-set tasks.
2. **Clinical-Aligned RL Optimization**: To facilitate free-form, open-ended diagnosis, we design a multi-component reward function encompassing localization accuracy, semantic correctness, and clinical safety. Driven by GRPO, this composite reward effectively aligns the VLM's generative capabilities with stringent clinical requirements, elevating its performance from rigid closed-ended metrics to reliable diagnostic narratives.
3. **Synthesis-Driven OOD Robustness**: We address the scarcity of rare clinical lesions by extending SynthSeg with pathological data augmentation, compelling the model to localize structural abnormalities during the RL

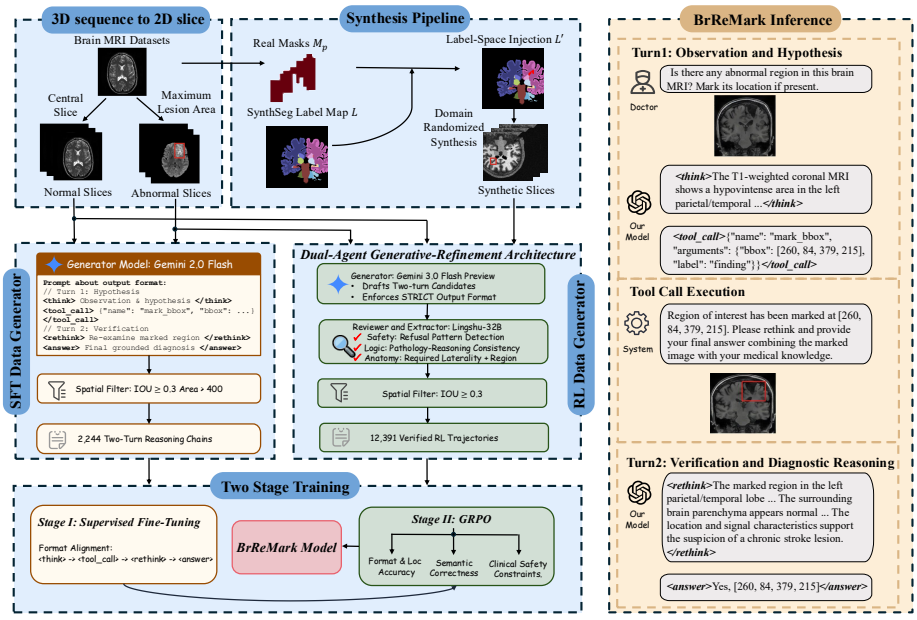

Fig. 1. Left: Data curation and training pipeline. Training data combines real abnormal slices, normal controls, and SynthSeg-generated slices. **Right:** BrReMark Inference pipeline. Given a brain MRI and query, the model first generates a hypothesis and invokes `mark_bbox` to localize suspicious regions (Turn 1), then re-examines the marked image to verify findings and produce a grounded diagnosis (Turn 2).

stage. This strategy enhances the model's robustness on unseen OOD cases, mirroring the complexities of real-world medical data.

4. **Comprehensive Evaluation**: We introduce a new two-turn interactive brain MRI dataset designed to realize BrReMark framework, representing a comprehensive integration of seven open-source datasets. We thoroughly validated BrReMark against internal and external benchmarks; it achieves superior results compared to other models within the same capacity class.

2 Methods

2.1 Overview

We formulate brain MRI diagnosis in the BReMark framework as a two-turn visual dialogue (Fig. 1). Given an image $\mathcal{I}$ and query q, the VLM π_θ first proposes a hypothesis h and a bounding box b (where $b = \texttt{null}$ for normal images): $o_1 = (h, b) \sim \pi_\theta(\cdot \mid \mathcal{I}, q)$. The box is then rendered to produce a marked image $\mathcal{I}_m$. In the second turn, the model verifies its initial findings using this augmented input

to output $o_2 = (v, y) \sim \pi_\theta(\cdot \mid \mathcal{I}, q, o_1, \mathcal{I}_m)$, yielding the verification reasoning v and the final diagnostic answer y.

During the training process, we employ a two-stage optimization strategy. Stage 1 utilizes SFT to instruct the VLM to adhere to the aforementioned two-turn diagnostic format. Subsequently, stage 2 leverages RL to ensure the model's reasoning and outputs align with human preferences.

2.2 Open Set Data Curation

To reflect real-world clinical challenges, which are inherently open-ended rather than closed-ended (e.g., multiple-choice), we curate a comprehensive open-set dataset for both the VLM's SFT and RL stages. We initially compile approximately 3,700 cases across five modalities (T1, T1-Gd, T2, FLAIR, DWI) and three anatomical planes (axial, coronal, sagittal). These are sourced from six abnormal datasets (BraTS-GLI/MEN/PED [3], UPENN-GBM [4], ATLAS [16], ISLES [11]) and one healthy control dataset (IXI [1]). During preprocessing, for each abnormal volumetric case, we extract a 480×480 2D slice containing the maximum lesion area based on the ground-truth mask, and generate its minimum bounding box for precise spatial localization. For normal cases from the IXI, we extract central 2D slices and explicitly set the bounding box to `null`.

Building upon this shared data pool, we first construct the SFT dataset. Since the SFT stage primarily requires learning format compliance, we prompt Gemini 2.0 Flash [8] with each image and its structured priors. This explicit guidance constrains the LLM to generate clinically grounded reasoning chains, directly yielding the two-turn interactive format incorporating `<think>`, `<tool_call>`, and `<rethink>` tags, as illustrated in Fig. 2(a).

RL stage necessitates a strictly structured QA format, shown in Fig. 2(b). To achieve this, we propose a **Dual-Agent Generative-Refinement Architecture** powered by the more advanced Gemini 3.0 Flash Preview [10]. Within this framework, a *Generator Agent* first produces an initial response encompassing the full cognitive sequence. Subsequently, a *Reviewer Agent* parses this output, extracting core clinical elements to populate the required template. RL data is further validated utilizing Lingshu-32B, which filters the structured outputs based on three criteria: (1) *Refusal Detection* (eliminating patterns like "I cannot"); (2) *Semantic Consistency* (ensuring logical coherence across reasoning phases); and (3) *Anatomical Standardization* (mandating precise spatial descriptions). After screening, we yield 12,391 high-quality structured RL samples.

2.3 Stage-wise Synthetic Pathology Injection for Generalization

To enhance OOD generalization, we introduce a synthetic pathology pipeline governed by a strict **Stage-Wise Data Routing** strategy. To avoid clinical hallucinations, synthetic data is excluded from SFT. It is used only in RL for spatial targeting, with diagnostic reasoning masked to ensure the model's foundational policy remains grounded in real-world physiology.

Specifically, to construct synthetic abnormal MRI volumes for the RL phase and improve diagnostic generalization to OOD data, we extract real lesion masks from the 6 aforementioned anomaly datasets and integrate them with healthy brain anatomical label maps provided by SynthSeg [6] (where each structure receives a distinct discrete label). Our synthesis pipeline comprises two stages:
Stage 1: Label-Space Pathology Injection. Given a healthy brain label map from SynthSeg, we inject real lesion masks into this label space. To determine valid implantation sites, we employ Euclidean Distance Transform (EDT) weighted sampling within internal brain tissues (e.g., white matter, cortex). This sampling strategy explicitly biases toward deep brain regions to prevent lesion boundaries from exceeding normal brain tissue margins.
Stage 2: Domain-Randomized Synthesis. We utilize SynthSeg generative model to synthesize images while maintaining lesion intensity distributions based on statistical analysis of real patient data. For a specific pathology, its intensity sampling prior is bounded by robust statistics (median $\pm$ 2$\times$MAD) derived from the real datasets. T1 tumor utilizes $\mu \in [88, 195]$, whereas DWI ischemic lesions employ $\mu \in [207, 277]$ to authentically reflect their characteristic hyperintensity.

2.4 Training Strategy

Stage 1: Supervised Fine Tuning. The primary objective is to teach the model the foundational two-turn paradigm and the structural syntax of the two-turn dialogue (e.g., `<think>`, `<tool_call>`). For a given image $\mathcal{I}$, query q, and target sequence $o^* = (o_1, o_2)$, we optimize the model parameters θ using standard autoregressive cross-entropy loss:

$$\mathcal{L}_{\text{SFT}} = -\log \pi_\theta(o_1 \mid \mathcal{I}, q) - \log \pi_\theta(o_2 \mid \mathcal{I}, q, o_1, \mathcal{I}_m). \tag{1}$$

By minimizing this loss, the model learns to first generate the hypothesis and bounding box (o_1), then verify its findings using the marked image $\mathcal{I}_m$ (o_2).
Stage 2: Reinforcement Learning. We design a multi-component reward function encompassing localization accuracy, semantic correctness, and clinical safety. For each input pair $(\mathcal{I}, q)$, we sample a group of G candidate responses. The composite reward for the i-th candidate is defined as $r_i = r_{\text{fmt}} + r_{\text{loc}} + r_{\text{llm}}$:

- **Format & Localization Accuracy:** To maintain structural integrity, $r_{\text{fmt}} \in [0, 0.4]$ evaluates reasoning tag completeness. For spatial precision, $r_{\text{loc}} \in [0, 0.5]$ computes bounding box accuracy via $\min(\text{IoU}, 0.5)$.
- **Semantic Correctness:** To robustly evaluate diagnostic narratives, $r_{\text{llm}} \in [0, 1.0]$ utilizes Lingshu-32B [24] as an LLM-as-a-judge [26]. It assesses the diagnostic validity of the generated text against the ground-truth (Fig. 2(b)).

Clinical Safety Constraints. To mitigate clinical hallucinations and guarantee reliability, we enforce three critical safety mechanisms acting as reward gatekeepers: (1) *Modality Gating:* Nullifies the semantic reward ($r_{\text{llm}} = 0$) if the MRI sequence is misidentified; (2) *Hallucination Penalty:* Sets the total reward

SFT Training Example (Normal)

Question: What do you observe in this brain MRI? Mark any suspicious regions.

Answer:
<think> T2 axial brain MRI. Symmetric ventricles, normal gray-white differentiation, no focal lesions or mass effect.</think>

<tool_call>{"bbox_2d": null, "label": "normal"}</tool_call>

<rethink>No abnormal signal. The brain parenchyma appears normal.</rethink>

<answer>No abnormalities detected in this brain MRI.</answer>

(a)

RL Training Example (Abnormal)

Question: Determine the most likely diagnosis from this axial diffusion-weighted brain MRI, with short evidence-based reasoning.
Structured Answer:
Bounding Box: [261, 261, 333, 308]
Reference Diagnosis:
- **Anatomy:** left parietal region
- **Signal:** hyperintensity on DWI
- **Diagnosis:** acute ischemic stroke
- **Reasoning:** Focal DWI hyperintensity in left parietal cortex, consistent with cytotoxic edema.
- **Differentials:** subacute stroke, encephalitis

(b)

Fig. 2. Training data formats for BrReMark. (a) SFT example demonstrating the two-turn reasoning format with <think> and <tool_call> tags for normal cases. (b) RL example showing structured answer format with ground-truth bounding box and reference diagnosis used for multi-component reward computation during GRPO training.

$r_i = 0$ if lesions are fabricated on healthy brain images; (3) *Synthetic Masking:* r_{llm} is excluded for synthetic samples to prevent artificial noise. Synthetic samples are also flagged in the prompt to distinguish them from real cases.

Optimization Objective. To update the model without requiring an external value network, GRPO normalizes the rewards within each group to compute the advantage $\hat{A}_i = (r_i - \mu_r)/\sigma_r$. The VLM policy π_θ is then optimized by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\rho_i\hat{A}_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\right) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right)\right], \quad (2)$$

where $\rho_i = \frac{\pi_\theta(o_i|\mathcal{I},q)}{\pi_{\text{ref}}(o_i|\mathcal{I},q)}$ represents the probability ratio of the current policy π_θ over the reference SFT model π_{ref}, ϵ limits the update step, and β controls the KL divergence penalty to prevent policy degradation.

3 Experiments and Results

3.1 Benchmark

We evaluate BrReMark across three clinical tasks: Anomaly Localization, Image Description, and Differential Diagnosis. Our curated dataset (Sec. 2.2) is partitioned 70%/10%/20% for training, validation, and testing. For description and diagnosis, we use expert reports from RadGenome-Brain_MRI [14] as ground truth for the ISLES and BraTS test subsets. For OOD evaluation, we employ NOVA [5], comprising 906 Eurorad scans across 281 rare pathologies. Excluded from all training phases, NOVA introduces substantial distribution shifts and long-tail anomalies to test the model's robustness to unseen diseases.

Table 1. Anomaly Localization Results. We evaluate with detection metrics (mAP at IoU 0.30/0.50, reported as percentages), number of true positives (TP_{30}), and number of false positives (FP_{30}). Bold numbers indicate the best results among models of comparable scale (≤30B).

Model	Our Benchmark				NOVA (OOD)			
	mAP_{30}	mAP_{50}	TP_{30}	$FP_{30}\downarrow$	mAP_{30}	mAP_{50}	TP_{30}	$FP_{30}\downarrow$
General Domain								
Qwen2.5-VL-72B	55.16	25.56	1323/2384	624	37.02	21.25	397/1068	525
LLaVA-v1.5-13B	9.31	3.36	222/2384	**140**	9.55	4.12	102/1068	**140**
InternVL3.5-8B	21.18	10.63	645/2384	2106	11.40	3.75	125/1068	919
Qwen2.5-VL-7B	11.50	3.72	274/2384	1157	21.23	9.77	227/1068	651
Medical Domain								
Lingshu-32B	4.19	1.24	155/2384	2285	6.56	1.77	96/1068	1122
HuatuoGPT-V-7B	14.14	3.89	337/2384	647	15.07	4.49	161/1068	617
MedVLM-R1	4.43	0.57	108/2384	1946	5.28	0.88	58/1068	831
Lingshu-7B	4.96	0.74	118/2384	1782	5.24	0.84	56/1068	836
Thinking with Images								
GRIT	15.73	6.27	353/2384	935	14.18	2.27	168/1068	742
BrReMark (Ours)	**64.93**	**37.54**	**1548/2384**	812	**33.05**	**13.30**	**353/1068**	285
BrReMark w/o synthesis	60.32	35.00	1438/2384	911	29.87	8.82	319/1068	388
BrReMark w/o r_{loc}	51.01	21.98	1216/2384	1161	23.25	6.93	251/1068	513
BrReMark w/o r_{llm}	62.78	36.58	1509/2384	852	30.81	11.52	329/1068	310
BrReMark w/o r_{fmt}	58.93	33.10	1405/2384	969	29.03	9.08	310/1068	459

3.2 Experimental Setup

Baselines and Implementation. We compare BrReMark against representative VLMs across three paradigms: (1) *General Domain*: Qwen2.5-VL-72B [2], InternVL3.5-8B [28], and Qwen2.5-VL-7B [2]; (2) *Medical Domain*: Lingshu-32B [24], MedVLM-R1 [19], HuatuoGPT-V-7B [7], and Lingshu-7B [24]; and (3) *Thinking with Images*: GRIT [9]. Lingshu-7B serves as both our foundation model and primary baseline. Training is conducted on 4×A100 80GB GPUs using verl [21]: SFT for 8 epochs ($lr = 3e-6$, batch size 16), followed by GRPO for 2 epochs ($lr = 1.5e-6$, $\beta = 0.001$, G=6). Training completes within 24 hours.
Evaluation Protocol. For localization and description, we adopt NOVA's evaluation metrics [5]. For diagnosis, we design an LLM-as-Judge framework where GPT-4o [12] evaluates three dimensions: *Diagnostic Accuracy* (semantic match with ground truth), *Reasoning Quality* (logical chain from imaging to diagnosis), and *Safety* (clinical appropriateness). All scores are normalized to 0–100%. Notably, our benchmark prompts provide only images and questions, whereas NOVA additionally incorporates clinical history.

3.3 Results

Superior Open-Set (ID) Performance. Despite being compact (7B), BrReMark excels across localization, description, and diagnosis in ID scenarios (Table 1 and 2). In anomaly localization, it achieves 64.93 mAP_{30}, outperforming

Table 2. Image Description and Differential Diagnosis on Our Benchmark and NOVA. Description quality is evaluated by METEOR, Clinical F1, and Modality F1. Diagnosis quality is evaluated by Diagnostic Accuracy, Reasoning Quality, and Safety. All metric values are reported as percentages.

Model	Description						Diagnosis					
	Our Benchmark			NOVA (OOD)			Our Benchmark			NOVA (OOD)		
	MET	Clin	Mod	MET	Clin	Mod	Diag	Reas	Safe	Diag	Reas	Safe
General Domain												
Qwen2.5-VL-72B	19.30	19.84	44.74	16.13	15.86	41.70	10.71	53.83	46.24	20.36	42.33	61.09
LLaVA-v1.5-13B	21.88	14.84	20.69	14.30	8.40	21.47	35.64	35.05	61.80	6.72	10.68	**48.30**
InternVL3.5-8B	20.52	14.86	30.95	17.18	15.51	57.49	9.73	27.34	28.89	11.20	26.04	40.85
Qwen2.5-VL-7B	18.67	18.18	41.83	14.50	14.32	37.84	7.15	42.23	43.15	6.28	**29.35**	30.23
Medical Domain												
Lingshu-32B	22.21	23.57	46.33	18.43	19.18	56.34	45.07	48.30	56.85	14.68	28.70	48.84
HuatuoGPT-V-7B	19.55	16.00	22.99	17.99	13.84	52.63	5.45	43.63	63.90	6.68	16.17	36.26
MedVLM-R1	19.51	12.63	18.74	15.69	13.19	30.51	10.95	28.26	30.61	10.14	15.09	23.02
Lingshu-7B	17.06	21.47	42.72	17.26	15.84	54.36	35.43	48.25	61.64	11.09	21.03	41.76
Thinking with Images												
GRIT	19.66	20.80	25.69	15.74	13.34	35.47	20.20	26.62	33.30	6.80	12.52	32.23
BrReMark (Ours)	**23.64**	21.57	**47.68**	**18.04**	**17.02**	**61.30**	**45.26**	**60.30**	**66.27**	10.43	26.32	43.07
BrReMark w/o synthesis	23.16	**23.44**	42.46	17.61	16.18	57.90	44.70	59.58	64.21	**11.53**	25.19	43.71
BrReMark w/o r_{loc}	23.38	23.04	45.65	17.65	15.04	58.41	44.33	58.03	63.18	8.83	24.29	34.05
BrReMark w/o r_{llm}	22.31	21.60	45.69	17.31	14.89	59.29	40.62	59.68	59.53	7.95	24.78	33.39
BrReMark w/o r_{fmt}	23.57	21.48	42.91	17.44	15.27	59.22	44.35	59.62	63.16	8.90	24.11	35.65

72B Qwen2.5-VL by 9.77. Precise grounding validates two-turn paradigm and provides explicit ROI for clinicians. For image description, BrReMark reaches a 21.57 Clinical F1, surpassing healthcare-adapted HuatuoGPT-V-7B (16.00) and trailing only Lingshu-32B. In clinical diagnosis, it achieves a leading Reasoning Quality score of 60.30. Powered by two-turn dataset, enhanced reasoning yields a state-of-the-art diagnostic accuracy of 45.26 (9.83 better than Lingshu-7B).

Hallucination Suppression. Compared to Lingshu-7B, BrReMark achieves this through: 1) Hallucination Penalty reduces detection false positives (FP) from 1782 to 812; 2) Modality Gating improves description Modality F1 from 42.72% to 47.68%. They benefit downstream diagnosis, where safety rises from 61.64% to 66.27%, with accuracy 45.26%, surpassing larger Lingshu-32B (45.07%).

Generalization under Distribution Shift. BrReMark demonstrates exceptional OOD generalization, primarily driven by the use of synthetic anomaly data during reinforcement learning. In anomaly detection, our 7B model dramatically increases mAP_{30} to 33.05 compared to Lingshu-32B (6.56), trailing only 72B Qwen2.5-VL due to parameter gap. For image description, it exhibits advantages against similar-sized models, improving the Clinical F1 (Clin) score from 13.84 (HuatuoGPT-V-7B) to 17.02. Although our synthetic data focuses on spatial localization rather than diagnostic semantics, it still yields improvements over the base Lingshu-7B, with the Safety score rising from 41.76 to 43.07.

Ablation Study. We conduct ablation studies (Tables 1 and 2) to validate key components. **(1)** The localization reward (r_{loc}) is crucial for spatial grounding;

removing it causes localization performance on our benchmark to drop precipitously from 64.93 to 51.01. **(2)** Synthetic anomaly data is pivotal for OOD robustness, as excluding it degrades OOD localization from 33.05 to 29.87. **(3)** The semantic correctness reward (r_{llm}) is indispensable for text-heavy description and diagnosis tasks; omitting it significantly deteriorates diagnostic accuracy from 45.26 to 40.62. **(4)** The reasoning format reward (r_{fmt}) is vital for guiding the model to mimic human-like reasoning processes; discarding it diminishes overall performance, reducing diagnostic accuracy from 45.26 to 44.35.

4 Conclusion

We present BrReMark, the first grounded reasoning framework for open-ended brain MRI diagnosis. Through a hypothesis-mark-verification cognitive chain, BrReMark unifies anomaly detection, description, and diagnosis within a single interactive paradigm. Our multi-component reward function enables RL training on open-ended outputs, achieving substantial improvements on BrReMark-Bench (mAP_{50}: 0.74%→37.54%) and strong generalization on NOVA. Future work would consider to direct 3D volumetric reasoning and multi-modal integration (e.g., CT, other MRI sequences), with prospective clinical validation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ixi dataset. https://brain-development.org/ixi-dataset/ (2023), accessed: 2023-02-15
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
4. Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., Shukla, G., Rudie, J., Flores Santamaria, N., Fathi Kazerooni, A., Pati, S., et al.: Multi-parametric magnetic resonance imaging (mpmri) scans for de novo glioblastoma (gbm) patients from the university of pennsylvania health system (upenn-gbm). The Cancer Imaging Archive **10** (2021)
5. Bercea, C., Li, J., Raffler, P., Riedel, E.O., Schmitzer, L., Kurz, A., Bitzer, F., Roßmüller, P., Canisius, J., Beyrle, M., et al.: Nova: A benchmark for rare anomaly localization and clinical reasoning in brain mri. Advances in Neural Information Processing Systems **38** (2026)
6. Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., et al.: Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. Medical image analysis **86**, 102789 (2023)

7. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Cai, Z., Ji, K., Wan, X., et al.: Towards injecting medical visual knowledge into multimodal llms at scale. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 7346–7370 (2024)
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. Fan, Y., He, X., Yang, D., Zheng, K., Kuo, C.C., Zheng, Y., Guan, X., Wang, X.: Grit: Teaching mllms to think with images. Advances in Neural Information Processing Systems **38**, 116522–116543 (2026)
10. Google DeepMind: Gemini 3: The next generation of multimodal ai. https://deepmind.google/technologies/gemini/ (2025), accessed: 2026-02-27
11. Hernandez Petzsche, M.R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., et al.: Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific data **9**(1), 762 (2022)
12. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
13. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5**(1), 1–10 (2018)
14. Lei, J., Zhang, X., Wu, C., Dai, L., Zhang, Y., Zhang, Y., Wang, Y., Xie, W., Li, Y.: Interpretable brain mri report generation anchored by lesion topography. IEEE Journal of Biomedical and Health Informatics (2025)
15. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
16. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data **9**(1), 320 (2022)
17. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1650–1654. IEEE (2021)
18. Liu, L., Yang, X., Lei, J., Shen, Y., Wang, J., Wei, P., Chu, Z., Qin, Z., Ren, K.: A survey on medical large language models: Technology, application, trustworthiness, and future directions. arXiv preprint arXiv:2406.03712 (2024)
19. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–347. Springer (2025)
20. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)

21. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. In: Proceedings of the Twentieth European Conference on Computer Systems. pp. 1279–1297 (2025)
22. Su, Z., Li, L., Song, M., Hao, Y., Yang, Z., Zhang, J., Chen, G., Gu, J., Li, J., Qu, X., et al.: Openthinkimg: Learning to think with images via visual tool reinforcement learning. arXiv preprint arXiv:2505.08617 (2025)
23. Wang, P., Zhang, H., He, Z., Peng, Z., Yuan, Y.: Ftspl: enhancing brain analysis with fmri-text synergistic prompt learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 564–574. Springer (2024)
24. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
25. Zhang, Y., Gao, J., Tan, Z., Zhou, L., Ding, K., Zhou, M., Zhang, S., Wang, D.: Data-centric foundation models in computational healthcare: A survey. ACM Computing Surveys **58**(11), 1–35 (2026)
26. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023)
27. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
28. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)