TOWARDS PERCEPTION-CENTRIC MULTIMODAL LLM FOR MULTI-SEQUENCE BREAST MRI DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have demonstrated remarkable promise in computer-assisted diagnosis, yet their clinical efficacy remains severely constrained in multi-sequence breast MRI diagnosis. We observe that existing models can achieve better diagnostic conclusions supplied with verified perceptual findings, yet their performance degrades drastically when operating directly on raw multi-sequence MR images, even under explicit clinical guidance. This discrepancy indicates that the primary bottleneck stems from visual perception. We attribute this deficiency to two primary obstacles, specifically clinically unguided perceptual exploration alongside weak multi-sequence perceptual capability. Current MLLMs fail to organize visual interpretation along established clinical perception logic, overlooking subtle yet diagnostically critical findings. Moreover, extracted perceptual cues are weakly grounded across MRI sequences, propagating perceptual errors into downstream diagnostic reasoning. To this end, we propose PCBreast, a perception-centric MLLM for multi-sequence breast MRI diagnosis. We construct MMBreast, a comprehensive multi-center, multi-sequence breast MRI dataset paired with perceptual Chain-of-Thought (CoT) trajectories, enabling the model to internalize clinically aligned perceptual exploration. Furthermore, we introduce Perception-Sensitive Policy Optimization (PSPO), which incorporates perception-grounded rubric reward and visual-sensitive credit assignment to incentivize faithful perceptual grounding and penalize spurious visual findings. Extensive experiments demonstrate that PCBreast consistently outperforms existing MLLMs, achieving impressive improvements in diagnostic accuracy. Our code and model will be publicly available at `https://anonymous.4open.science/w/PCBreast-2D2A/`.

1 INTRODUCTION

Breast cancer remains one of the leading causes of cancer-related incidence and mortality among women worldwide (Bray et al., 2024; Sun et al., 2017). In modern clinical workflows, multi-sequence MRI serves as the gold-standard modality for early detection, therapeutic monitoring, and preoperative staging (Mann et al., 2019; Pinker et al., 2017). By synthesizing anatomical structures from T1-weighted imaging (T1), edema and fluid distribution from T2-weighted imaging (T2), microstructural cellularity from Diffusion-Weighted Imaging (DWI) and Apparent Diffusion Coefficient (ADC) maps, as well as vascular hemodynamics from Dynamic Contrast-Enhanced (DCE) MRI, multi-sequence breast MRI delivers indispensable diagnostic insights (Kuhl, 2007; Luo et al., 2025). Concurrently, the emergence of MLLMs has sparked widespread interest in automated medical image interpretation (Moor et al., 2023a; Zhang et al., 2024). Endowed with extensive biomedical knowledge, these models show considerable promise in tasks ranging from visual question answering to computer-aided diagnosis (Wu et al., 2025; Nie et al., 2026).

Despite these advances, deploying MLLMs directly to multi-sequence breast MRI diagnosis remains fundamentally challenging (Hager et al., 2024). As illustrated in Figure 1, while supplying models with verified perceptual findings substantially boosts diagnostic accuracy, operating directly on raw MR scans yields severely depressed performance that barely benefits from explicit clinical guidance (Xu et al., 2026; Jiang et al., 2025). Moreover, introducing visual corruptions into the input MRI sequences causes negligible performance degradation, revealing that existing models fail to ground

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.

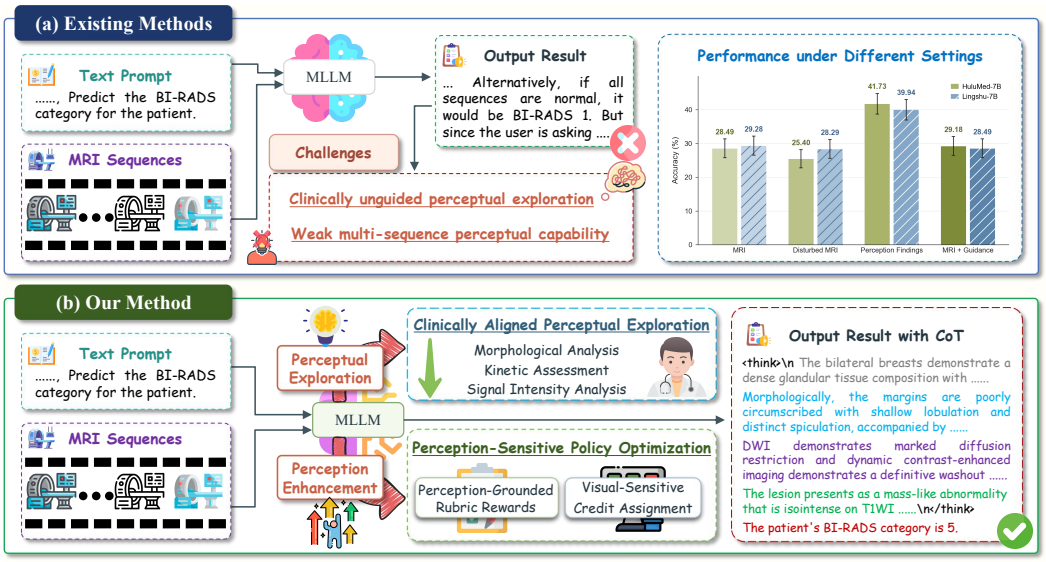


Figure 1: **Comparison between existing methods and PCBreast.** (a) Current MLLMs suffer from clinically unguided exploration and weak multi-sequence perceptual capability, resulting in spurious reasoning and misdiagnoses. (b) Our framework addresses this via guided perceptual exploration and perception-sensitive policy optimization to reinforce faithful visual perception.

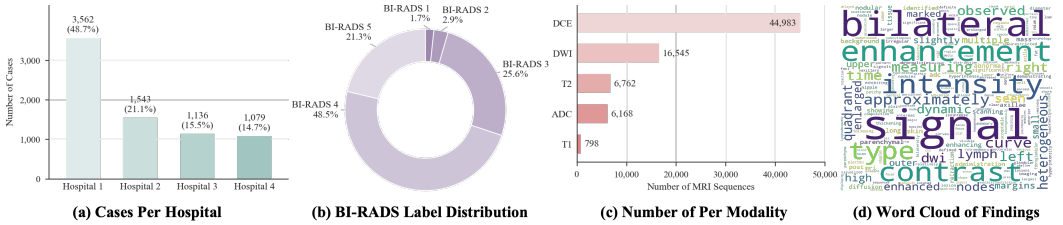
their decisions on authentic imaging evidence. Together, these empirical findings converge on an unequivocal conclusion: the primary bottleneck resides in visual perception.

We attribute this perceptual deficiency to two core obstacles including clinically unguided perceptual exploration and weak multi-sequence perceptual capability. Clinically unguided perceptual exploration arises because current models lack clinical guidance priors. Expert radiologists systematically interpret scans along established clinical perceptual guidance, existing MLLMs conduct disorganized perceptual exploration, overlooking subtle yet diagnostically decisive abnormalities (Chen et al., 2024; Sellergren et al., 2025). In addition, weak multi-sequence perceptual capability constrains the reliability to perceive visual evidence. Even when explicitly instructed on what features to assess, current solutions fail to resolve intricate imaging patterns, generating spurious visual cues that inevitably cascade into diagnostic errors (Li et al., 2023; Yuan et al., 2026).

To overcome these challenges, we propose PCBreast, a perception-centric MLLM designed specifically for multi-sequence breast MRI diagnosis. To eliminate disorganized visual exploration, we construct MMBreast, a comprehensive multi-center, multi-sequence breast MRI dataset paired with perceptual CoT trajectories. Grounded on MMBreast, our clinically aligned perceptual exploration framework internalizes standard clinical diagnostic logic directly into the model's reasoning trajectory, systematically guiding PCBreast through sequential morphological, kinetic, and signal-intensity assessments before drawing diagnostic conclusions. Moreover, to overcome weak multi-sequence perception and eliminate unreliable visual cues, we introduce PSPO, incorporating perception-grounded rubric reward paired with visual-sensitive credit assignment. By evaluating perceptual fidelity against clinical ground truth and explicitly attributing optimization feedback to visual-sensitive tokens, PSPO enforces faithful visual perception and penalizes spurious radiological findings. Extensive experimental evaluations demonstrate that PCBreast consistently outperforms existing general-purpose and medical MLLMs.

The primary contributions of this work are summarized as follows:

- We identify the critical visual perception bottleneck in multi-sequence breast MRI diagnosis and propose PCBreast, a perception-centric MLLM designed to overcome clinically unguided exploration and weak multi-sequence perception capability.
- We construct MMBreast, a multi-center multi-sequence breast MRI dataset paired with perceptual exploration CoT trajectories, enabling the model to internalize clinical perceptual logic across morphological, kinetic, and signal-intensity dimensions.



(a) Cases Per Hospital **(b) BI-RADS Label Distribution** **(c) Number of Per Modality** **(d) Word Cloud of Findings**

Figure 2: **Overview of MMBreast dataset.** Key statistics of MMBreast dataset, summarizing case breakdowns across centers, BI-RADS categories, sequence volumes and radiological descriptors.

- We introduce PSPO, which incorporates perception-grounded rubric reward to evaluate perceptual fidelity and visual-sensitive credit assignment to steer gradient updates toward visual-sensitive tokens, enforcing reliable visual perception.
- Extensive evaluations demonstrate that PCBreast achieves state-of-the-art diagnostic accuracy, outperforming existing general-purpose and specialized medical MLLMs.

2 Related Work

Clinical breast MRI interpretation strictly follows multi-sequence protocols governed by standardized BI-RADS criteria, motivating computer-aided systems to evolve from handcrafted radiomic features to multi-parametric deep representation fusion (Kuhl, 2007; Mann et al., 2015; Pinker et al., 2017; Truhn et al., 2019; Zhang et al., 2023b). Concurrently, medical multimodal large language models have exhibited remarkable capabilities in automated report generation and clinical reasoning across diverse modalities (Moor et al., 2023a; Singhal et al., 2023; Zhou et al., 2024; Bannur et al., 2024). Recent endeavors progressively advance toward volumetric structural modeling and preference optimization paradigms to align diagnostic outputs with expert clinical standards (Hung et al., 2024; Singhal et al., 2025). A full version is available in the Appendix A.

3 MMBreast Dataset

3.1 Dataset Overview

To establish a comprehensive benchmark for multi-sequence breast MRI diagnosis, we construct MMBreast from four independent medical institutions. As illustrated in Figure 2, MMBreast aggregates over 7,000 examinations across centers. Across these studies, we assemble over 70,000 sequences, establishing an extensive foundation for multi-sequence breast MRI diagnosis. Each examination in MMBreast is paired with verified reference diagnoses and detailed radiological findings structured according to the ACR BI-RADS lexicon (Rao et al., 2016). As shown in Figure 2, the diagnostic distribution spans categories from BI-RADS 1 to 5, faithfully mirroring routine clinical practice and presenting nuanced categorical boundaries that demand rigorous multi-parametric reasoning. Notably, examinations categorized as BI-RADS 0 and BI-RADS 6 are excluded to ensure definitive categorical evaluation and focus exclusively on primary diagnostic characterization. Furthermore, the paired radiological text encapsulates granular visual evidence, capturing morphological characteristics, kinetic enhancement curves, and sequence-specific signal intensities.

3.2 Clinical Perception Logic

In clinical practice, multi-sequence breast MRI interpretation strictly adheres to standardized diagnostic pathways defined by the ACR BI-RADS lexicon (Rao et al., 2016). Expert radiologists do not infer malignancy directly from holistic scans; instead, they conduct a structured diagnostic progression across complementary imaging sequences (Baltzer et al., 2013). To eliminate disorganized visual scanning and ground the perceptual process, we formalize this diagnostic workflow into three progressive perceptual dimensions:

- **Morphological Analysis:** Radiologists inspect anatomical landmarks and lesion geometry, wherein critical visual indicators include mass shape, margin definition, and non-mass enhancement distribution patterns.

- **Kinetic Assessment:** Using multi-phase DCE MRI, clinicians assess the microvascular perfusion and vessel permeability of suspected lesions. By comparing pre-contrast and delayed phases, they categorize time-intensity curves into persistent, plateau, or rapid washout kinetics, which indicate the degree of tumor angiogenesis.
- **Signal-Intensity Analysis:** Hypercellularity and fluid dynamics are cross-referenced across non-contrast sequences. Markedly restricted diffusion on paired DWI and ADC maps indicates dense tumor cell proliferation, while T2 signal characteristics differentiate fluid-rich benign cysts and identify peritumoral edema associated with invasive phenotypes.

This clinical perception logic establishes an explicit mapping between diverse MRI modalities and definitive diagnostic criteria. By decomposing complex visual evidence into sequentially coupled morphological, kinetic, and intensity evaluations, the framework mitigates perceptual confusion.

3.3 Perceptual Exploration CoT Construction

To instantiate the structured clinical logic into verifiable reasoning trajectories, we construct high-quality CoT data as shown in Figure 3. Given the multi-sequence scans, ground-truth BI-RADS labels, and radiological findings, Gemini 3.8 Flash (Team et al., 2023) is prompted to synthesize diagnostic rationales that systematically step through morphological, kinetic, and signal-intensity dimensions. To prevent perceptual hallucination, each candidate trajectory undergoes automated verification by Qwen 3.7 Max (Yang et al., 2025) against authentic clinical findings, followed by iterative regeneration and expert radiological review. This process ultimately yields rigorous reasoning paths that faithfully connect multi-sequence visual evidence to definitive diagnostic conclusions.

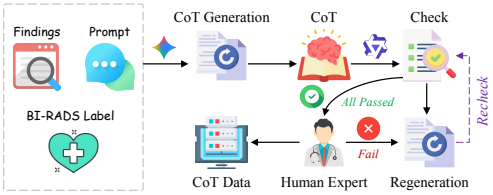


Figure 3: **The CoT construction workflow.** This workflow synthesizes clinically aligned perceptual exploration trajectories.

4 Method

4.1 Problem Formulation

Given a comprehensive breast MRI examination, the multimodal input consists of a multi-sequence image collection denoted by $\mathcal{X} = \{X_m\}_{m=1}^{M}$, where each X_m represents an image acquisition corresponding to a specific clinical protocol. In addition to the multi-parametric scans, the model receives a clinical task prompt T_{in}. Our objective is to learn a multimodal foundation model parameterized by Θ to generate a clinically structured response sequence $Y = (y_1, y_2, \ldots, y_N)$.

Our framework structures the output into an explicit CoT reasoning trajectory $R = (y_1, \ldots, y_K)$ followed by a definitive diagnostic verdict $C = (y_{K+1}, \ldots, y_N)$. The reasoning sequence R is structured to progress through three distinct perceptual dimensions: morphological analysis R_{morph}, kinetic assessment R_{kin}, and signal-intensity evaluation R_{sig}. Formally, the sequence likelihood is factorized autoregressively as:

$$P_\Theta(Y \mid \mathcal{X}, T_{\text{in}}) = \prod_{t=1}^{N} P_\Theta(y_t \mid y_{<t}, \mathcal{X}, T_{\text{in}}). \tag{1}$$

Under this formulation, multi-sequence visual features extracted by the vision encoder are projected into the textual token space via the visual adapter, enabling the autoregressive backbone to ground its clinical reasoning along verified perceptual evidence before concluding the final decision C.

4.2 Clinically Aligned Perceptual Exploration

To resolve the challenge of clinically unguided perceptual exploration, we establishe an explicit bridge between raw multi-sequence MRI and clinical perception logic. We accomplish this through multi-sequence visual representation alignment and clinical diagnostic trajectory internalization.

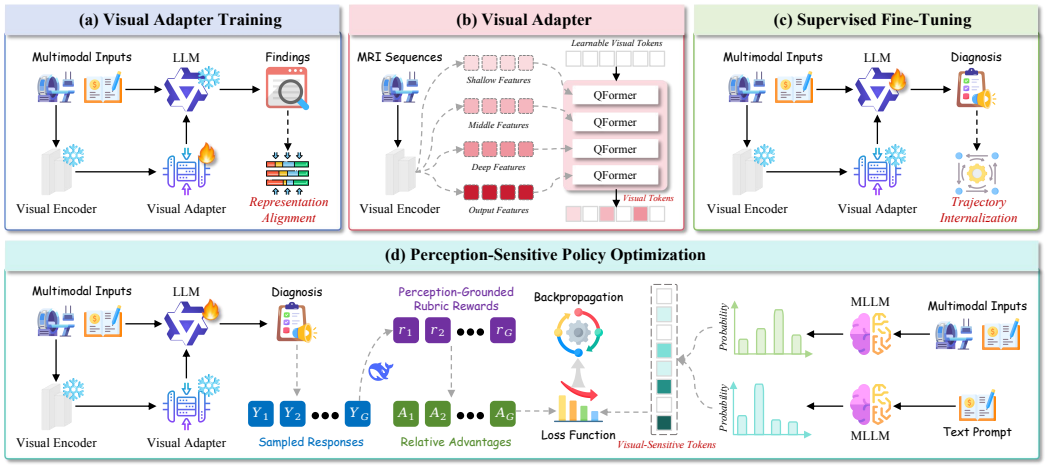


Figure 4: **Overview of PCBreast.** (a) Visual Adapter Training: Aligns multi-sequence MRI features into language space. (b) Visual Adapter: Distills multi-level representations into a compact set of visual tokens. (c) SFT: Internalizes perceptual trajectories into the model. (d) PSPO: Enforces reliable perception via perception-grounded rubric reward and visual-sensitive credit assignment.

Multi-Sequence Visual Representation Alignment. Multi-sequence breast MRI entails distinct physical properties. Directly prompting an LLM to reason over raw visual features results in unsatisfactory results. To bridge this multimodal representation gap, we first train the visual adapter on paired radiological findings, as illustrated in Figure 4(a).

Specifically, the composite multi-sequence volume $\mathcal{X}$ is processed by the vision encoder $\mathcal{E}_v$ to capture visual representations $V = \mathcal{E}_v(\mathcal{X})$. As shown in Figure 4(b), the visual adapter $\mathcal{P}_v$ takes the hierarchical features extracted from $\mathcal{E}_v$ and employs multi-layer Q-Formers to project them into a compact set of l visual tokens, which are then mapped into the language feature space:

$$H_v = \mathcal{P}_v(V). \tag{2}$$

Given an instruction prompt T_{find} querying the detailed findings, the model is trained to autoregressively reconstruct the authentic radiological findings sequence $F = (f_1, f_2, \ldots, f_{|F|})$. During this stage, both the vision encoder $\mathcal{E}_v$ and the language backbone $\mathcal{M}_{\text{LLM}}$ remain frozen. The negative log-likelihood objective is used:

$$\mathcal{L}_{\text{align}}(\Theta) = -\sum_{t=1}^{|F|} \log P_\Theta\big(f_t \mid f_{i<t}, \mathcal{X}, T_{\text{find}}\big). \tag{3}$$

Clinical Diagnostic Trajectory Internalization. To eliminate disorganized visual exploration, we internalize the standardized clinical logic into the model by fine-tuning it along perceptual trajectories, as depicted in Figure 4(c). Given the multi-sequence volume $\mathcal{X}$ and the task prompt T_{in}, the target sequence is structured into a sequential trajectory $Y = (R_{\text{morph}}, R_{\text{kin}}, R_{\text{sig}}, C)$, progressing through morphological analysis, kinetic assessment, signal-intensity evaluation, and the definitive BI-RADS verdict C. We optimize the model using the supervised fine-tuning objective:

$$\mathcal{L}_{\text{sft}}(\Theta) = -\sum_{t=1}^{|Y|} \log P_\Theta\big(y_t \mid y_{<t}, \mathcal{X}, T_{\text{in}}\big). \tag{4}$$

This causal formulation compels the model to ground each perceptual judgment on preceding clinical evidence, internalizing the diagnostic progression and ensuring that the final verdict C is supported by the perceptual findings.

4.3 Perception-Sensitive Policy Optimization

We introduce PSPO coupling perception-grounded rubric reward with visual-sensitive credit assignment to reinforce authentic visual perception as depicted in Figure 4(d).

Perception-Grounded Rubric Reward. To prevent reward hacking and ensure diagnostic predictions stem from genuine visual evidence, we formulate a composite reward balancing categorical correctness and fine-grained perceptual fidelity. For each input $(\mathcal{X}, T_{\text{in}})$, the policy samples a group of G candidate trajectories $\{Y_i\}_{i=1}^{G} = \{(R_i, C_i)\}_{i=1}^{G}$ under Group Relative Policy Optimization (GRPO). The composite reward r_i for trajectory Y_i is defined as:

$$r_i = \lambda_{\text{acc}} \cdot r_{\text{acc},i} + \lambda_{\text{rubric}} \cdot r_{\text{rubric}}(R_i, R^*), \tag{5}$$

where $r_{\text{acc},i}$ denotes the categorical accuracy reward comparing the predicted BI-RADS category C_i against clinical ground truth C^*, and $\lambda_{\text{acc}}, \lambda_{\text{rubric}}$ denote trade-off weights.

To evaluate visual perceptual fidelity, an automated LLM judge $\mathcal{M}_{\text{judge}}$ assesses the generated reasoning narrative R_i against the verified clinical reference R^*. The evaluation systematically examines three clinical perceptual dimensions $\mathcal{D} = \{\text{morph}, \text{kin}, \text{sig}\}$, which correspond to lesion geometry and margin morphology, dynamic contrast-enhancement perfusion kinetics, and multi-parametric signal intensities, respectively. Formally, the rubric reward is defined as:

$$r_{\text{rubric}}(R_i, R^*) = \frac{1}{9} \sum_{d \in \mathcal{D}} s_d, \quad \text{with} \quad s_d = \mathcal{M}_{\text{judge}}(d \mid R_i, R^*), \tag{6}$$

where $s_d \in \{0, 1, 2, 3\}$ denotes the alignment score, with higher values indicating greater factual consistency with the clinical reference.

Visual-Sensitive Credit Assignment. Standard GRPO assigns a uniform trajectory advantage across all response tokens. However, in breast MRI diagnosis, only a subset of tokens are sensitive to genuine visual evidence, while the remainder represents generic linguistic templates. Uniform policy gradient updates inadvertently reward ungrounded text, encouraging spurious shortcuts.

To accurately identify visual-sensitive tokens, we introduce visual-sensitive credit assignment. For each response token y_t, we quantify visual sensitivity Δ_t as the log-likelihood discrepancy between the multimodal forward pass and an ablated unimodal pass without visual inputs:

$$\Delta_t = \log \pi_\Theta(y_t \mid y_{<t}, \mathcal{X}, T_{\text{in}}) - \log \pi_\Theta(y_t \mid y_{<t}, T_{\text{in}}). \tag{7}$$

We construct the evidence set S by filtering the top-k proportion with positive sensitivity, assigning unnormalized weights $w_t = \min(1 + \gamma\Delta_t, w_{\max})$ for $t \in S$ and $w_t = 1$ otherwise, where γ scales sensitivity and $w_{\max}$ caps extreme values.

To maintain optimization stability and preserve the gradient scale of standard GRPO, we enforce a scale-preserving normalization:

$$\tilde{w}_t = \frac{w_t}{\sum_{j \in \text{valid}} w_j}. \tag{8}$$

We then integrate $\tilde{w}_t$ into the policy gradient surrogate:

$$\mathcal{L}_t^{\text{PSPO}} = -\tilde{w}_t \cdot \min\left(\rho_t A_i,\ \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon) A_i\right) \tag{9}$$

Here, $\rho_t = \pi_\Theta(y_t \mid y_{<t}, \mathcal{X}, T_{\text{in}}) / \pi_{\text{old}}(y_t \mid y_{<t}, \mathcal{X}, T_{\text{in}})$ represents the token-level importance sampling ratio, A_i is the standardized advantage computed across the candidate trajectory group $\{Y_i\}_{i=1}^{G}$, and ϵ specifies the clipping margin.

5 Experiments

5.1 Experiment Setup

Implementation Details. We implement PCBreast using the SWIFT (Zhao et al., 2024) framework on eight NVIDIA GPUs with Qwen3.5-2B backbone (Yang et al., 2025). All MRI inputs are resized to 224×224. In visual adapter training, the visual encoder and LLM are frozen while the compressor is trained for 1,000 steps using AdamW with learning rate 5×10^{-4} and cosine scheduler at a global batch size of 128. SFT is set for 500 steps with a global batch size of 128, a learning rate of 1.5×10^{-4} with LoRA. PSPO is trained for 400 steps with a learning rate of 5×10^{-6}, a KL penalty $\beta = 0.04$. For each prompt, the policy samples $G = 16$ candidate trajectories at temperature 1.0, weighting diagnostic accuracy and perception-grounded rubric reward at $\lambda_{\text{acc}} = 1.0$ and $\lambda_{\text{rubric}} =$

Table 1: **Diagnostic performance on MMBreast benchmark.** All metrics are evaluated against ground truth. Best results are highlighted in **bold**, and second-best results are underlined.

Method	Params	Overall Metrics			Multi-Center Accuracy			
		Accuracy	*Macro F1*	*Weighted F1*	*Hospital-1*	*Hospital-2*	*Hospital-3*	*Hospital-4*
General Foundation Models								
InternVL3.5	1B	0.1365	0.1015	0.1628	0.2110	0.1423	0.1429	0.0530
InternVL3.5	2B	0.2918	0.1641	0.2404	0.4128	0.2659	0.3524	0.2119
InternVL3.5	4B	0.3337	0.1528	0.2454	0.4312	0.3614	0.3143	0.1921
InternVL3.5	8B	0.3137	0.1627	0.2277	0.3578	0.2622	0.5286	0.1656
Qwen3.5	0.8B	0.2311	0.1527	0.2000	0.1651	0.1461	0.4619	0.2583
Qwen3.5	2B	0.2430	0.1644	0.2106	0.2477	0.1966	0.4000	0.1854
Qwen3.5	4B	0.2560	0.1194	0.1565	0.1651	0.1517	0.6000	0.2119
Qwen3.5	9B	0.2171	0.1301	0.1764	0.1835	0.1180	0.4619	0.2517
Medical Foundation Models								
HuatuoGPT-Vision	7B	0.0757	0.0528	0.0564	0.1560	0.0693	0.0286	0.1060
ClinFusion	8B	0.2082	0.1609	0.2235	0.2936	0.2378	0.1048	0.1854
Hulu-Med	4B	0.2629	0.1239	0.1965	0.3761	0.3034	0.1524	0.1921
Hulu-Med	7B	0.2849	0.1303	0.1839	0.3945	0.2959	0.2619	0.1987
Hulu-Med	14B	0.2888	0.1214	0.1768	0.4128	0.2903	0.2952	0.1854
Hulu-Med	32B	0.2789	0.1655	0.2165	0.2294	0.1873	0.5857	0.2119
Lingshu	7B	0.2928	0.1092	0.1545	0.4404	0.2678	0.3619	0.1788
MedGemma	4B	0.2789	0.1390	0.1942	0.3670	0.1610	0.5905	0.1987
PCBreast	2B	**0.5189**	**0.3333**	**0.4966**	**0.4771**	**0.5225**	**0.6571**	**0.3444**

3.0, respectively. Visual-sensitive credit assignment is parameterized with a top-k filtering ratio of 0.2, sensitivity scale $\gamma = 1.0$, and clipping threshold $w_{\max} = 2.5$. The token budget of visual adapter l is set to 1024. We use MMBreast dataset that is composed of 6,316 samples for training and 1,002 for testing. To be specific, 6,316, 4,816 and 1,500 cases are used for visual adapter training, SFT and reinforcement learning, respectively.

Baselines. To comprehensively assess the efficacy of PCBreast, we benchmark our method against an extensive suite of competitive multimodal foundation models, categorized into two groups: *General Foundation Models*: We evaluate leading general-purpose vision-language architectures across diverse model capacities, including the InternVL3.5 (Wang et al., 2025) family and the Qwen3.5 series (Yang et al., 2025). For those general foundation models, the preprocessing pipeline of MRI sequences is followed by Lingshu (Xu et al., 2026). *Medical Foundation Models*: To examine the effect of domain-specific adaptation, we benchmark against specialized medical MLLMs, including ClinFusion (Yuan et al., 2026), HuatuoGPT-Vision (Chen et al., 2024), Lingshu (Xu et al., 2026), MedGemma (Sellergren et al., 2025), and Hulu-Med (Jiang et al., 2025).

5.2 Main Results

Comparison with General Foundation Models. General foundation models exhibit severely constrained performance on multi-sequence breast MRI diagnosis. As shown in Table 1, both the InternVL3.5 and Qwen3.5 families achieve suboptimal results across varied model capacities. Notably, the best-performing general model, InternVL3.5-4B, achieves an overall accuracy of only 33.37% and a Macro F1 of 15.28%. In sharp contrast, our PCBreast achieves an accuracy of 51.89% and a Macro F1 of 33.33%, outperforming the strongest general baseline by 18.52% in accuracy and more than doubling the Macro F1 score. In addition, PCBreast demonstrates robust diagnostic superiority across all four clinical sites.

Comparison with Medical Foundation Models. Medical foundation models likewise fail to demonstrate a competitive edge over their general counterparts on multi-sequence breast MRI diagnosis. As indicated in Table 1, Lingshu-7B achieves the highest accuracy among all medical baselines at 29.28%, while Hulu-Med-32B obtains the top Macro F1 of only 16.55%. Notably, scaling model capacity across the Hulu-Med series from 4B to 32B yields marginal gains and saturates below 29% accuracy. PCBreast establishes new state-of-the-art diagnostic performance, surpassing the strongest medical baseline Lingshu-7B in accuracy and Macro F1 score.

Table 2: **Evaluation of perceptual fidelity.** We report BERTScore-F1 (BERT.) and LLM-judge score (LLM.). Best results are highlighted in **bold**, and second-best results are underlined.

Model	Params	Overall		Morphological		Kinetic		Signal Intensity	
		BERT.↑	LLM.↑	BERT.↑	LLM.↑	BERT.↑	LLM.↑	BERT.↑	LLM.↑
InternVL3.5	4B	0.8717	0.3134	0.8760	0.2581	0.8798	0.2805	0.8584	**0.4111**
Qwen3.5	4B	0.8667	0.2880	0.8690	0.1922	0.8717	0.2849	0.8590	0.3992
Hulu-Med	7B	0.8737	0.3054	0.8702	0.2576	0.8862	0.2907	0.8642	0.3722
Lingshu	7B	0.8722	0.3281	0.8724	0.2856	0.8818	0.3143	0.8617	0.3915
SFT (ours)	2B	0.8826	0.3762	0.8778	0.3186	0.8898	0.4239	0.8806	0.3850
PCBreast	2B	**0.8839**	**0.3869**	**0.8790**	**0.3237**	**0.8906**	**0.4330**	**0.8824**	0.4058

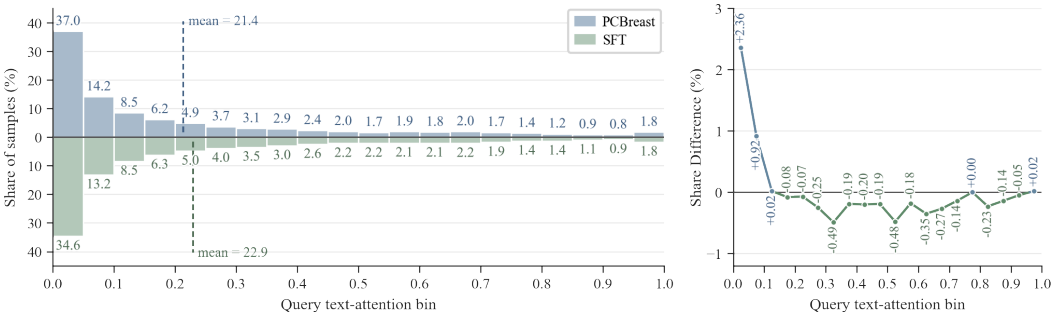


Figure 5: **Attention analysis of query text-attention distributions.** The left panel illustrating sample proportions across attention bins and the right panel plotting their differential curve.

5.3 Empirical Analysis

Perception Fidelity. As reported in Table 2, our method consistently surpasses competitive domain-specific medical baselines across all clinical perceptual dimensions. Specifically, the LLM-judge metric outputs a normalized score within the range of 0 to 1, where higher scores reflect superior perceptual precision and factual alignment against the ground-truth perceptual findings from the MMBreast dataset. We employ Qwen 3.7 Max as the judge LLM. For PCBreast and the SFT baseline, perceptual assessments are parsed directly from their generated CoT reasoning trajectories, whereas other baselines are evaluated by querying them with explicit perceptual task prompts. In particular, our model demonstrates substantial improvements in both BERTScore-F1 and LLM-judge metrics across morphological, kinetic, and signal-intensity evaluations, achieving marked advantages in kinetic assessment over Hulu-Med and Lingshu. Furthermore, introducing PSPO on top of SFT achieves the highest overall alignment score, verifying that perception-sensitive policy updates effectively steer the reasoning trajectory toward faithful visual evidence.

Attention Analysis of Query. PCBreast shifts the query text-attention distribution toward lower-density bins compared to the SFT baseline, with the differential curve revealing a marked reduction in excessive text-attention proportions, as shown in Figure 5. This shift indicates that our solution effectively suppresses over-reliance on textual prompt shortcuts, ensuring that the diagnostic process is predominantly driven by multi-sequence visual evidence.

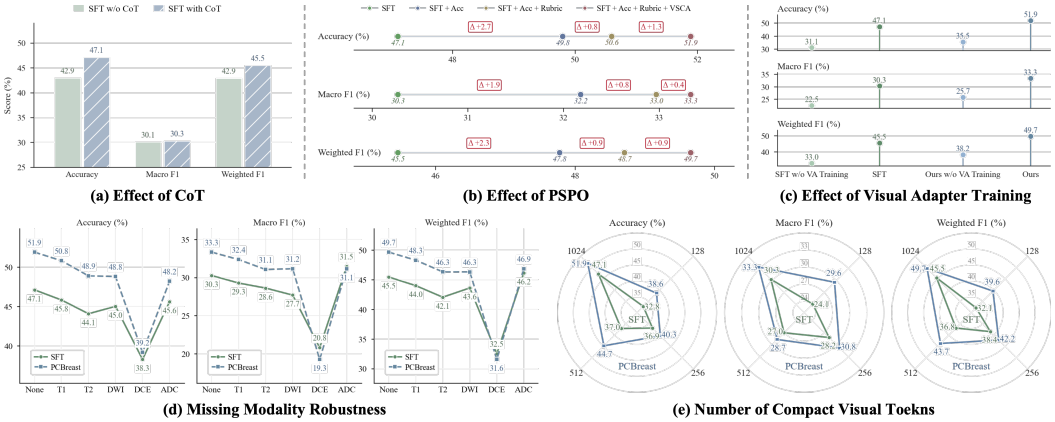
Visualization of Visual-Sensitive Tokens. Visual-sensitive credit assignment accurately allocates higher sensitivity weights to evidence-aware radiological descriptors, while keeping generic linguistic connectors unweighted, as illustrated in Figure 6, . This fine-grained token attribution confirms that policy optimization gradients are steered toward authentic multi-sequence visual cues, thereby preventing the model from exploiting superficial textual templates.

5.4 Ablation Study

Effect of Perceptual CoT. Incorporating perceptual CoT yields substantial performance gains across all evaluation metrics compared with standard direct-prediction fine-tuning, as depicted in Figure 7(a). Structuring the diagnostic trajectory through sequential morphological, kinetic, and

The primary consideration is whether the combination of irregular morphology, restricted diffusion, and plateau kinetics warrants tissue sampling. The lesion presents as a solitary mass with an irregular shape, indistinct margins, and heterogeneous internal enhancement. On functional sequences, it demonstrates high signal on DWI with corresponding low signal on the ADC map, confirming restricted diffusion, and exhibits a plateau (Type II) time-signal intensity curve. Morphologically, the irregular shape and indistinct margins raise concern for a malignant process, such as invasive carcinoma, rather than a benign fibroadenoma. The functional parameters, specifically the restricted diffusion and plateau kinetics, further support an active, cellular neoplastic process. While a benign lesion such as a sclerosing adenosis or a hyalinized fibroadenoma could be considered in the differential, the irregular margins, heterogeneous enhancement, and restricted diffusion argue against a typical benign entity. The presence of a plateau kinetic curve rather than a washout curve keeps the lesion within the suspicious range, precluding a definitive BI-RADS 5 classification, while the irregular morphology and diffusion restriction clearly exceed the threshold for a probably benign BI-RADS 3 assessment. These features favor BI-RADS 4.

Figure 6: **Visualization of visual-sensitive tokens from the perceptual exploration CoT.** Color-highlighted words reflect higher visual sensitivity.



(a) Effect of CoT (b) Effect of PSPO (c) Effect of Visual Adapter Training

(d) Missing Modality Robustness (e) Number of Compact Visual Toekns

Figure 7: **Ablation studies on key components of PCBreast.** These studies evaluate the individual impacts of perceptual exploration trajectories, PSPO optimization components, visual adapter training, modality-missing robustness, and visual token budgets.

signal-intensity assessments effectively prevents hasty conclusions, guiding the model to ground categorical predictions on progressive radiological reasoning.

Effectiveness of PSPO. We validate the individual contributions of our PSPO components in Figure 7(b). Integrating the perception-grounded rubric reward substantially elevates diagnostic fidelity, and further coupling it with visual-sensitive credit assignment achieves optimal results by directing policy updates toward visual-sensitive tokens.

Impact of Visual Adapter Training. We evaluate the necessity of visual adapter training in Figure 7(c). Completely discarding this process leads to marked performance drops, confirming that visual representation alignment with structured findings bridges the modality representation gap and provides a critical foundation for diagnostic reasoning.

Robustness of Missing Modality. We evaluate diagnostic resilience under various single-sequence missing scenarios in Figure 7(d). Although excluding functional sequences like DCE or DWI predictably hampers performance, PCBreast consistently preserves superior diagnostic accuracy and exhibits significantly milder degradation than the baseline SFT across all settings.

Selection of Visual Token Budgets. We investigate model sensitivity to the number of visual tokens ranging from 128 to 1024, as shown in Figure 7(e). The diagnostic performance across all metrics improves consistently as the token capacity increases.

6 Conclusion

We identify that the primary bottleneck hindering medical MLLMs in multi-sequence breast MRI diagnosis stems from visual perception. To address clinically unguided exploration and weak multi-sequence perception, we propose PCBreast, a perception-centric framework. We construct MMBreast, a multi-center dataset paired with perceptual exploration trajectories that is applied to internalize sequential morphological, kinetic, and signal-intensity diagnostic logic into the reasoning process. Furthermore, we present perception-sensitive policy optimization, which couples perception-grounded rubric rewards with visual-sensitive credit assignment to steer policy updates toward visual-sensitive tokens and eliminate spurious shortcuts. The experiments demonstrate that PCBreast outperforms existing models across diverse clinical sites.

AI USE STATEMENT

In this work, we used generative AI tools for synthesizing CoT data, automated evaluation and code generation. We have not used generative AI tools for other tasks. All LLM-generated data and code were verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses the data from 4 hospitals. The study protocol was approved by the Ethics Committee of `[Hospital Names]`, and registered under Clinical Trial No. `[XXXX]`. The identifying information has been anonymized for double-blind review and will be disclosed in the final version upon publication. All participants were informed of the purpose and procedure of the study, the use of their responses for research, and their right to withdraw, and informed consent was obtained before participation. Participants were compensated in accordance with the `[Hospital Name]` guidelines. We did not collect or release directly identifying information, and the study results are reported only in aggregate. Any participant-derived data released with this work will be de-identified and shared only to the extent permitted by participant consent, applicable regulations, and institutional policies.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, model architecture, training and inference procedures, and evaluation protocol in the main paper. To facilitate reproducibility and future research, we will publicly release the source code, pretrained model weights, and the constructed internal dataset upon publication, applicable regulations, and regional policies.

REFERENCES

Pascal AT Baltzer, Matthias Dietzel, and Werner A Kaiser. A simple and robust classification tree for differentiation between benign and malignant lesions in mr-mammography. *European radiology*, 23(8):2051–2060, 2013.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Neha Bhooshan, Maryellen L Giger, Sanaz A Jansen, Hui Li, Li Lan, and Gillian M Newstead. Cancerous breast lesions on dynamic contrast-enhanced mr images: computerized characterization for image-based prognostic markers. *Radiology*, 254(3):680–690, 2010.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74 (3):229–263, 2024.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, 2024. URL https://arxiv.org/abs/2406.19280.

Monique D Dorrius, Marijke C Jansen-van der Weide, Peter MA van Ooijen, Ruud M Pijnappel, and Matthijs Oudkerk. Computer-aided detection in breast mri: a systematic review and meta-analysis. *European radiology*, 21(8):1600–1608, 2011.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

Alex Ling Yu Hung, Haoxin Zheng, Kai Zhao, Xiaoxi Du, Kaifeng Pang, Qi Miao, Steven S Raman, Demetri Terzopoulos, and Kyunghyun Sung. Csam: A 2.5 d cross-slice attention module for anisotropic volumetric medical image segmentation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5911–5920. IEEE, 2024.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Christiane Kuhl. The current status of breast mr imaging part i. choice of technique, image interpretation, diagnostic accuracy, and transfer to clinical practice. *Radiology*, 244(2):356–378, 2007.

Sibel Kul, Aysegul Cansu, Etem Alhan, Hasan Dinc, Gurbuz Gunes, and Abdulkadir Reis. Contribution of diffusion-weighted imaging to dynamic contrast-enhanced mri in the characterization of breast tumors. *American Journal of Roentgenology*, 196(1):210–217, 2011.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564, 2023.

Luyang Luo, Mingxiang Wu, Mei Li, Yi Xin, Qiong Wang, Varut Vardhanabhuti, Winnie CW Chu, Zhenhui Li, Juan Zhou, Pranav Rajpurkar, et al. A large model for non-invasive and personalized management of breast cancer from multiparametric mri. *Nature communications*, 16(1):3647, 2025.

Ritse M Mann, Corinne Balleyguier, Pascal A Baltzer, Ulrich Bick, Catherine Colin, Eleanor Cornford, Andrew Evans, Eva Fallenberg, Gabor Forrai, Michael H Fuchsjäger, et al. Breast mri: Eusobi recommendations for women's information. *European radiology*, 25(12):3669–3678, 2015.

Ritse M Mann, Nariya Cho, and Linda Moy. Breast mri: state of the art. *Radiology*, 292(3):520–536, 2019.

Ellen B Mendelson, Marcela Böhm-Vélez, Wendie A Berg, GJ Whitman, MI Feldman, H Madjar, et al. Acr bi-rads® ultrasound. *ACR BI-RADS® atlas, breast imaging reporting and data system*, 2013, 2013.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023a.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023b.

Yuxiang Nie, Sunan He, Yequan Bie, Yihui Wang, Zhixuan Chen, Shu Yang, Zhiyuan Cai, Linshan Wu, Hongmei Wang, Xi Wang, et al. An explainable biomedical foundation model via large-scale concept-enhanced vision–language pretraining. *Nature Biomedical Engineering*, pp. 1–14, 2026.

Savannah C Partridge, Zheng Zhang, David C Newitt, Jessica E Gibbs, Thomas L Chenevert, Mark A Rosen, Patrick J Bolan, Helga S Marques, Justin Romanoff, Lisa Cimino, et al. Diffusion-weighted mri findings predict pathologic response in neoadjuvant treatment of breast cancer: the acrin 6698 multicenter trial. *Radiology*, 289(3):618–627, 2018.

K Pinker, H Bickel, TH Helbich, S Gruber, P Dubsky, U Pluschnig, M Rudas, Z Bago-Horvath, M Weber, S Trattnig, et al. Combined contrast-enhanced magnetic resonance and diffusion-weighted imaging reading adapted to the "breast imaging reporting and data system" for multiparametric 3-t imaging of breast lesions. *European radiology*, 23(7):1791–1802, 2013.

Katja Pinker, Thomas H Helbich, and Elizabeth A Morris. The potential of multiparametric mri of the breast. *The British journal of radiology*, 90(1069):20160715, 2017.

Ajay Aroor Rao, Jennifer Feneis, Chloe Lalonde, and Haydee Ojeda-Fournier. A pictorial review of changes in the bi-rads fifth edition. *Radiographics*, 36(3):623–639, 2016.

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.

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.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

David Allen Spak, John S Plaxco, Laura Santiago, Michael J Dryden, and Burc E Dogan. Bi-rads® fifth edition: A summary of changes. *Diagnostic and interventional imaging*, 98(3):179–190, 2017.

Yi-Sheng Sun, Zhao Zhao, Zhang-Nv Yang, Fang Xu, Hang-Jing Lu, Zhi-Yong Zhu, Wen Shi, Jianmin Jiang, Ping-Ping Yao, and Han-Ping Zhu. Risk factors and preventions of breast cancer. *International journal of biological sciences*, 13(11):1387, 2017.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Daniel Truhn, Simone Schrading, Christoph Haarburger, Hannah Schneider, Dorit Merhof, and Christiane Kuhl. Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast mri. *Radiology*, 290(2):290–297, 2019.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Teresa C Williams, Wendy B DeMartini, Savannah C Partridge, Sue Peacock, and Constance D Lehman. Breast mr imaging: computer-aided evaluation program for discriminating benign from malignant lesions. *Radiology*, 244(1):94–103, 2007.

Reiko Woodhams, Saadallah Ramadan, Peter Stanwell, Satoko Sakamoto, Hirofumi Hata, Masanori Ozaki, Shinichi Kan, and Yusuke Inoue. Diffusion-weighted imaging of the breast: principles and clinical applications. *Radiographics*, 31(4):1059–1084, 2011.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hangjie Yuan, Yichen Qian, Zhiwei Tang, Xianzhe Xu, Lirong Wu, Sicheng Yang, Jinwang Wang, Pengju Wang, Zhitao Zeng, Yizeng Han, et al. Clinfusion: A vision-centric multimodal llm system for holistic medical understanding. *arXiv preprint arXiv:2607.24743*, 2026.

Jun Zhang, Ashirbani Saha, Zhe Zhu, and Maciej A Mazurowski. Hierarchical convolutional neural networks for segmentation of breast tumors in mri with application to radiogenomics. *IEEE transactions on medical imaging*, 38(2):435–447, 2018.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature medicine*, 30(11):3129–3141, 2024.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3):6, 2023a.

Tianyu Zhang, Luyi Han, Anna D'Angelo, Xin Wang, Yuan Gao, Chunyao Lu, Jonas Teuwen, Regina Beets-Tan, Tao Tan, and Ritse Mann. Synthesis of contrast-enhanced breast mri using t1-and multi-b-value dwi-based hierarchical fusion network with attention mechanism. In *International conference on medical image computing and computer-assisted intervention*, pp. 79–88. Springer, 2023b.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Huan Zheng, Yucheng Zhou, Tianyi Yan, Dubing Chen, Hongbo Lu, Wenlong Liao, Tao He, Pai Peng, and Jianbing Shen. Clinical cognition alignment for gastrointestinal diagnosis with multimodal llms. *arXiv preprint arXiv:2603.20698*, 2026.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649, 2024.

Yucheng Zhou, Huan Zheng, Dubing Chen, Hongji Yang, Wencheng Han, and Jianbing Shen. From medical llms to versatile medical agents: A comprehensive survey. 2025. URL https://openreview.net/pdf?id=75M55jtwj6.

A Related Work

A.1 Breast MRI Diagnosis

Breast MRI serves as the gold-standard imaging modality for high-risk screening, preoperative staging, and response monitoring in breast oncology (Kuhl, 2007; Mann et al., 2015; Partridge et al., 2018). In clinical practice, reliable diagnosis strictly adheres to standardized guidelines, such as the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) lexicon (Mendelson et al., 2013; Spak et al., 2017). To differentiate malignant lesions from benign findings, radiologists perform holistic interpretations across multi-sequence MRI protocols (Pinker et al., 2013; Kul et al., 2011). Specifically, T1 characterizes anatomical landmarks and baseline signal distributions; T2 evaluates tissue hydration, cyst structures, and peritumoral edema; DWI paired with ADC maps quantifies restricted water diffusion indicative of hypercellularity; and DCE MRI depicts microvascular architecture through hemodynamic kinetic curves (Woodhams et al., 2011). Synthesizing these sequences enables a systematic diagnostic trajectory across lesion morphology, contrast enhancement kinetics, and multi-parametric signal intensities (Pinker et al., 2017).

To alleviate the workload of clinicians, extensive computer-aided diagnosis frameworks have been developed for multi-sequence breast MRI interpretation (Dorrius et al., 2011; Williams et al., 2007). Early methods focused primarily on handcrafted radiomic features, extracting quantitative descriptors of tumor shape, texture, and kinetic curves to train traditional machine learning classifiers (Bhooshan et al., 2010). With the progression of deep learning, convolutional neural networks and vision transformers have been widely investigated for breast lesion detection, segmentation, and malignancy prediction (Truhn et al., 2019; Zhang et al., 2018). Subsequent advances have introduced multi-parametric fusion strategies, utilizing cross-attention mechanisms, hierarchical feature aggregation, and cross-sequence alignment modules to capture complementary representations across diverse MRI sequences (Zhang et al., 2023b).

A.2 Medical Multimodal Large Language Model

The remarkable success of general MLLMs has catalyzed rapid development in domain-specific foundation models for healthcare (Moor et al., 2023a; Singhal et al., 2023). By integrating visual representation encoders with autoregressive language backbones through cross-modal projectors, medical MLLMs exhibit versatile competencies spanning clinical visual question answering, automated diagnostic report generation, and multi-turn medical consultation (Zheng et al., 2026; Moor et al., 2023b; Zhou et al., 2025). Pioneering frameworks adapt general vision-language architectures to the biomedical domain via large-scale pretraining and clinical instruction tuning on paired medical image-text datasets (Zhang et al., 2023a). Representative systems, such as medical instruction-following models and specialized clinical dialogue agents, have demonstrated substantial diagnostic proficiency across modalities including chest radiographs, dermatological photographs, and histopathological whole-slide images (Zhou et al., 2024).

To address more complex clinical examinations, recent endeavors have progressively expanded medical MLLMs toward multi-image, multi-view, and volumetric imaging modalities (Bannur et al., 2024). Several contemporary approaches incorporate 3D patch tokenizers, cross-slice attention mechanisms, or spatial-temporal aggregators to preserve spatial continuity and volumetric context across consecutive image slices (Hung et al., 2024). In parallel, reinforcement learning from human feedback (RLHF) and preference optimization paradigms have been introduced to align model responses with expert clinical standards, thereby penalizing medically unsafe outputs and encouraging factually coherent reasoning trajectories (Singhal et al., 2025).

B Pseudo-code of PSPO

Algorithm 1 summarizes the training pipeline of PSPO. Throughout optimization, the visual components ($\mathcal{E}_v$ and $\mathcal{P}_v$) remain frozen to supply a fixed representation prefix $H_v = \mathcal{P}_v(\mathcal{E}_v(\mathcal{X}))$, while policy gradients exclusively update the language backbone parameters Θ.

Algorithm 1 PSPO Training Pipeline

Require: Dataset $\mathcal{D}$; frozen visual extractor $(\mathcal{E}_v, \mathcal{P}_v)$; policy π_Θ; reference π_{ref}; learning rate η.

Ensure: Optimized language backbone parameters Θ.

while training **do**
  Sample input $(\mathcal{X}, T_{\text{in}}) \sim \mathcal{D}$ and extract frozen visual prefix $H_v = \mathcal{P}_v(\mathcal{E}_v(\mathcal{X}))$
  Sample G trajectories $\{Y_i\}_{i=1}^{G} \sim \pi_\Theta(\cdot \mid H_v, T_{\text{in}})$, where $Y_i = (R_i, C_i)$
  Compute composite rewards r_i and group-relative advantages A_i via Eq. (5)
  for each trajectory $Y_i = (y_1, y_2, \ldots, y_N)$ **do**
    Compute visual sensitivity for each token:
      $\Delta_t \leftarrow \log \pi_\Theta(y_t \mid y_{<t}, H_v, T_{\text{in}}) - \log \pi_\Theta(y_t \mid y_{<t}, T_{\text{in}})$
    Select evidence set $S \leftarrow \text{TopK}(\{t \mid \Delta_t > 0\}, k)$
    Set unnormalized weights: $w_t \leftarrow \min(1 + \gamma\Delta_t, w_{\max})$ for $t \in S$, else $w_t \leftarrow 1$
    Compute normalized weights: $\tilde{w}_t \leftarrow w_t / \sum_{j \in \text{valid}} w_j$
    Calculate token surrogate losses: $\ell_t \leftarrow \tilde{w}_t \cdot \mathcal{L}_t^{\text{GRPO}} + \beta \cdot \mathcal{L}_t^{\text{KL}}$ via Eq. (9)
  end for
  Update trainable backbone parameters: $\Theta \leftarrow \Theta - \eta \nabla_\Theta \sum_{i=1}^{G} \sum_{t=1}^{N} \ell_t^{(i)}$
end while
return Optimized parameter set Θ

C Analysis of Visual-Sensitive Credit Assignment

This section analyzes the local effect of visual-sensitive credit assignment on token-level policy-gradient contributions based on the objective $\mathcal{L}_t^{\text{PSPO}}$ defined in Eq. (9).

Local Policy-Gradient Interpretation. To understand how visual-sensitive credit assignment shapes optimization dynamics, we examine the local gradient behavior around the current policy rollout. At the operational point where $\Theta = \Theta_{\text{old}}$, the token-level importance sampling ratio evaluates to unity, $\rho_{i,t}(\Theta_{\text{old}}) = 1$. Because this value falls strictly within the clipping boundary $1 \in (1 - \epsilon, 1 + \epsilon)$, the clipped surrogate objective in Eq. (9) locally coincides with its unclipped branch. Treating both the group-relative advantage A_i and the normalized sensitivity weight $\tilde{w}_{i,t}$ as stop-gradient values, the parameter gradient for token t is expressed as:

$$\nabla_\Theta \mathcal{L}_{i,t}^{\text{PSPO}}(\Theta)\big|_{\Theta=\Theta_{\text{old}}} = -\tilde{w}_{i,t} A_i \mathbf{g}_{i,t}, \tag{10}$$

where the standard score function vector $\mathbf{g}_{i,t}$ is given by:

$$\mathbf{g}_{i,t} = \nabla_\Theta \log \pi_\Theta \left(y_{i,t} \mid y_{i,<t}, H_{v,i}, T_i\right)\big|_{\Theta=\Theta_{\text{old}}}. \tag{11}$$

Summing across all valid response tokens $t \in \mathcal{V}_i$, the aggregated policy gradient for candidate trajectory i becomes:

$$\nabla_\Theta \sum_{t \in \mathcal{V}_i} \mathcal{L}_{i,t}^{\text{PSPO}}(\Theta)\Bigg|_{\Theta=\Theta_{\text{old}}} = -A_i \sum_{t \in \mathcal{V}_i} \tilde{w}_{i,t} \mathbf{g}_{i,t} = -\frac{A_i}{Z_i} \sum_{t \in \mathcal{V}_i} w_{i,t} \mathbf{g}_{i,t}, \tag{12}$$

where $Z_i = \sum_{j \in \mathcal{V}_i} w_{i,j}$ denotes the trajectory-specific normalization factor. Equation equation 12 formally demonstrates that PSPO directly modulates each token's gradient contribution through its unnormalized visual sensitivity coefficient $w_{i,t}$.

To quantify the differential update between evidence-bearing and generic text, consider a visual-sensitive token $t_s \in S_i$ and an unweighted linguistic token $t_u \notin S_i$ with $w_{i,t_u} = 1$. The ratio of

Table 3: **Fine-grained evaluation of perceptual fidelity on four different hospitals.** We report BERTScore-F1 (BERT.) and LLM-judge score (LLM.) across overall perception and three clinical perceptual dimensions, with all results reported as percentages (%).

Model	Params	Overall		Hospital-1		Hospital-2		Hospital-3		Hospital-4	
		BERT.	LLM.	BERT.	LLM.	BERT.	LLM.	BERT.	LLM.	BERT.	LLM.
Overall Perception											
InternVL3.5	4B	87.17	31.34	86.80	33.43	87.31	26.82	87.35	39.56	86.66	34.60
Qwen3.5-4B	4B	86.67	28.80	86.60	32.39	86.62	24.18	87.07	36.68	86.33	31.77
Hulu-Med	7B	87.37	30.54	87.33	29.81	87.46	29.51	87.34	33.87	87.13	30.17
Lingshu	7B	87.22	32.81	87.79	34.72	87.20	28.78	87.34	39.18	86.70	36.96
SFT (ours)	2B	88.26	37.62	87.38	32.41	88.60	36.77	88.61	46.80	87.24	31.88
PCBreast	2B	88.39	38.69	87.62	34.22	88.67	36.49	88.54	47.47	87.70	37.65
Morphological Analysis											
InternVL3.5	4B	87.60	25.81	87.62	25.55	87.67	25.27	87.60	27.05	87.33	26.19
Qwen3.5	4B	86.90	19.22	87.21	23.26	86.69	18.22	87.15	21.17	87.07	17.15
Hulu-Med	7B	87.02	25.76	87.46	25.52	86.73	25.43	87.46	27.76	87.11	24.36
Lingshu	7B	87.24	28.56	87.81	31.85	87.10	25.97	87.40	35.17	87.12	26.32
SFT(ours)	2B	87.78	31.86	88.30	26.06	87.60	32.57	88.00	35.76	87.73	28.23
PCBreast	2B	87.90	32.37	88.48	29.22	87.71	32.58	88.20	36.43	87.74	28.39
Kinetic Assessment											
InternVL3.5	4B	87.98	28.05	87.94	27.20	88.34	22.17	87.72	44.27	87.08	28.14
Qwen3.5	4B	87.17	28.49	87.58	26.98	87.29	22.90	87.21	41.78	86.43	31.22
Hulu-Med	7B	88.62	29.07	88.97	24.89	89.10	30.06	87.78	31.03	87.79	25.59
Lingshu	7B	88.18	31.43	89.58	30.33	88.29	27.68	87.94	38.85	87.17	35.97
SFT(ours)	2B	88.98	42.39	87.99	31.56	89.79	44.61	88.57	48.39	87.22	33.28
PCBreast	2B	89.06	43.30	88.07	29.34	89.82	44.33	88.42	49.41	87.79	40.31
Signal Intensity Analysis											
InternVL3.5	4B	85.84	41.11	85.00	45.42	85.75	34.29	86.70	48.69	85.58	49.98
Qwen3.5	4B	85.90	39.92	85.10	44.10	85.82	32.89	86.82	48.69	85.50	48.01
Hulu-Med	7B	86.42	37.22	85.83	37.67	86.39	33.50	86.73	42.99	86.50	41.01
Lingshu	7B	86.17	39.15	86.30	40.75	86.06	33.68	86.65	43.69	85.81	49.63
SFT(ours)	2B	88.06	38.50	85.88	38.06	88.41	31.64	89.31	57.96	86.81	35.21
PCBreast	2B	88.24	40.58	86.44	41.65	88.50	31.23	89.02	58.27	87.63	46.96

their corresponding gradient magnitudes satisfies:

$$\frac{\left\| \nabla_\Theta \mathcal{L}^{\text{PSPO}}_{i,t_s}(\Theta)\big|_{\Theta=\Theta_{\text{old}}} \right\|}{\left\| \nabla_\Theta \mathcal{L}^{\text{PSPO}}_{i,t_u}(\Theta)\big|_{\Theta=\Theta_{\text{old}}} \right\|} = \frac{\tilde{w}_{i,t_s}}{\tilde{w}_{i,t_u}} \frac{\|\mathbf{g}_{i,t_s}\|}{\|\mathbf{g}_{i,t_u}\|} = \min\left(1+\gamma\Delta_{i,t_s}, w_{\max}\right) \frac{\|\mathbf{g}_{i,t_s}\|}{\|\mathbf{g}_{i,t_u}\|}. \tag{13}$$

This theoretical relation proves that PSPO deterministically scales up gradient updates for tokens anchored on authentic visual evidence, effectively suppressing the reinforcement of superficial linguistic shortcuts and promoting faithful cross-modal perception.

D Detailed Perception Results

To thoroughly investigate how well MLLMs capture fine-grained radiological evidence, we present comprehensive evaluations of perceptual fidelity across individual clinical institutions in Table 3. The evaluation measures factual alignment between generated reasoning trajectories and verified radiological ground truth, decomposed into three standardized BI-RADS perceptual dimensions: morphological analysis, kinetic assessment, and signal intensity analysis.

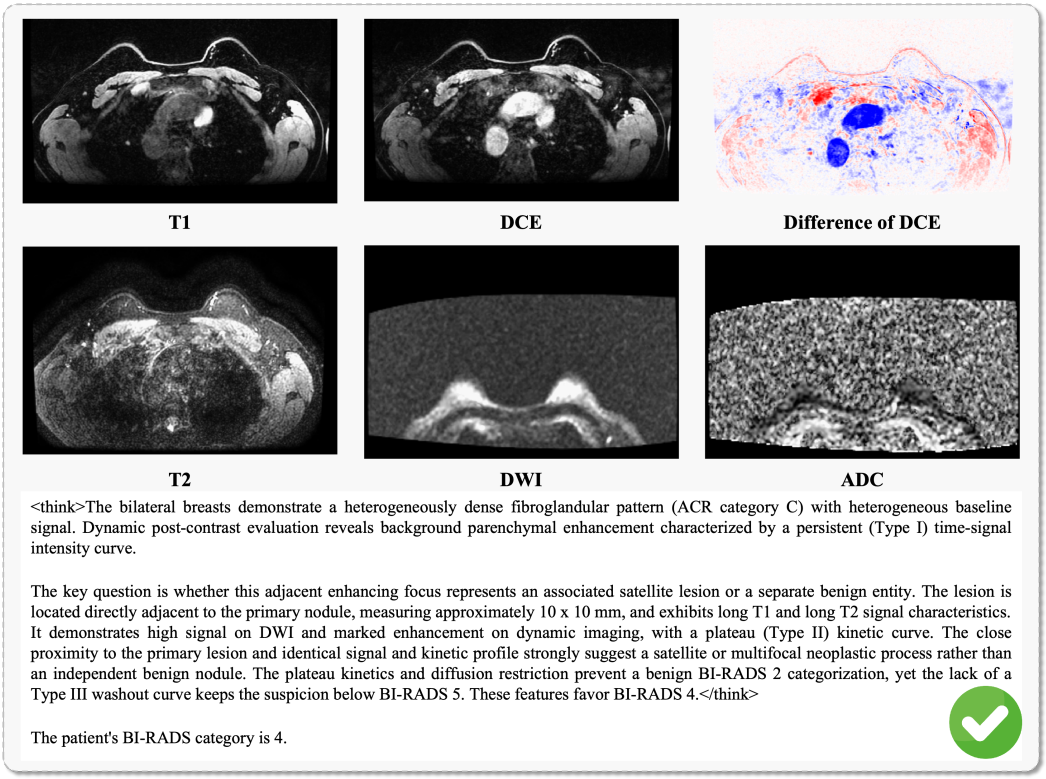


Figure 8: **Case study of our PCBreast.**

Superiority across Diverse Perceptual Dimensions. As evidenced by the empirical results, PCBreast consistently outperforms all general-purpose and domain-specialized baselines across every clinical perceptual sub-domain, despite operating at a significantly more compact parameter scale. In morphological analysis, PCBreast achieves an overall LLM-judge score of 32.37% and a BERTScore-F1 of 87.90%, substantially surpassing leading generalist models such as InternVL3.5-4B and specialized medical models such as Hulu-Med-7B. The advantage becomes even more prominent in kinetic assessment, where models must track subtle hemodynamic changes across multi-phase DCE sequences. In this challenging setting, PCBreast attains an overall LLM-judge score of 43.30%, establishing a remarkable lead over Hulu-Med-7B and Lingshu-7B. These substantial improvements confirm that our clinically aligned trajectory supervision enables the model to accurately capture lesion vascularity and dynamic enhancement patterns rather than reciting generic medical templates.

Consistent Multi-Center Generalization. A critical vulnerability of clinical vision-language systems is the performance drop under cross-center distribution shifts, which typically arise from varying scanner manufacturers, acquisition parameters, and magnetic field strengths. Table 3 demonstrates that PCBreast maintains high perceptual fidelity across all four independent clinical cohorts. For instance, in kinetic assessment, PCBreast achieves 44.33% in Hospital-2 and 49.41% in Hospital-3, while general baselines like InternVL3.5 and Qwen3.5 experience severe degradation, dropping to approximately 22% in Hospital-2. Similarly, in signal-intensity analysis, PCBreast reliably interprets diffusion restriction and T2 fluid dynamics across sites, achieving 58.27% in Hospital-3 and 46.96% in Hospital-4. This cross-site consistency indicates that grounding perceptual reasoning in clinical logic significantly enhances model robustness against domain discrepancy.

Effectiveness of PSPO. Comparing the SFT baseline against the complete PCBreast framework highlights the essential role of PSPO. While SFT already offers a solid foundation through perceptual exploration trajectory internalization, applying PSPO provides consistent improvements across all perceptual dimensions, elevating the overall LLM-judge score from 37.62% to 38.69% and signal intensity fidelity from 38.50% to 40.58%. The perception-grounded rubric reward together with visual-sensitive credit assignment successfully directs gradient updates toward authentic visual to-

kens. Consequently, the model learns to penalize hallucinated descriptors and selectively reinforce clinically reliable evidence across complementary MRI sequences.

E Case Study

We show the case study of our proposed PCBreast in Figure 8. As illustrated, PCBreast successfully identifies the critical morphological, kinetic, and signal-intensity characteristics of the lesion. By systematically reasoning over these multi-sequence perceptual findings, the model correctly predicts the final BI-RADS category 4, demonstrating its strong and reliable diagnostic capabilities.

F Details of Perception-Grounded Rubric Reward

We show the details of the used prompt for perception-grounded rubric reward:

Prompt for Perception-Grounded Rubric Reward

You are a radiology quality-control reviewer. You compare two English breast-MRI reports:

- Candidate (A): a model-generated reasoning narrative.
- Reference (B): the ground-truth reasoning narrative derived from the clinical report.

Judge how faithfully A reflects the lesion-characterization facts stated in B across THREE dimensions:

1. **morphology** (0-3): shape, margins, spiculation, lobulation, boundary.
2. **kinetics** (0-3): dynamic enhancement curve type (persistent/type I, plateau/type II, washout/type III) and enhancement pattern.
3. **signal** (0-3): T1/T2/DWI/STIR signal characteristics and ADC values.

Scoring rubric (per dimension):

- 3 = fully aligned: A states the same facts as B (same curve type, same signal pattern, same morphology), no missing key assertion, no fabrication.
- 2 = partially aligned: main facts match, but some lesions/values differ or a secondary assertion is missing/added.
- 1 = vague/tangential: A only mentions the dimension in generic terms ("kinetics assessed", "signals reviewed") without concrete matching assertions.
- 0 = contradictory/fabricated: A states facts opposite to B (e.g. B says persistent, A says washout), or A fabricates specific features that B does not contain (e.g. restricted diffusion / irregular margins absent from B).

Rules:

- Fabrication (A claims a specific feature absent from B) is scored 0 for that dimension; a mere omission is at most 2.
- WHOLE-TEXT COVERAGE: You MUST examine every paragraph and sentence of A, including its final paragraphs, before scoring. Do not skip trailing content. If A contains ANY lesion or feature assertion (e.g. an extra lesion, irregular margins, restricted diffusion, a curve type, an ADC value) that does not appear in B, that is fabrication and the affected dimension is 0, even when the rest of A closely matches B.
- You MUST quote the exact phrases from A and B that support the score as evidence. Do not introduce terms that appear in neither text.
- Output ONLY a JSON object (no markdown):

```
{
"morphology": <0-3>, "kinetics": <0-3>, "signal": <0-3>,
 "evidence": {"morphology": "A says '...', B says '...'",
 "kinetics": "...", "signal": "..."},
 "summary": "one-sentence overall assessment"
 }
```

Candidate (A - model-generated): candidate
Reference (B - ground truth): reference
Follow the rubric in the system prompt and return the JSON only.