ArogyaSutra: A Multi-Agent Framework for Multimodal Medical Reasoning in Indic Languages

Tanmoy Kanti Halder[1,3*], **Akash Ghosh**[1*], **Subhadip Baidya**[2], **Arijit Roy**[1], **Sriparna Saha**[1]
[1]Indian Institute of Technology Patna
[2]Indian Institute of Technology Kanpur
[3]Prasannadeb Women's College
{tanmoy.zx, subhadipbaidya23, akashghosh.ag90}@gmail.com, {sriparna, arijitroy}@iitp.ac.in

Abstract

Multimodal Large Language Models (MLLMs) have shown promising reasoning capabilities in general domains, yet their performance remains limited in specialized settings such as healthcare, especially in multilingual and low-resource scenarios. This gap is critical in regions like rural India, where patients often express complex medical queries in native Indic languages and rely on multimodal inputs such as medical images. Existing English-centric MLLMs struggle to support such use cases, limiting equitable access to AI-driven healthcare assistance. To address this challenge, we introduce **ArogyaBodha**, a large-scale multilingual multimodal medical question-answer dataset constructed from eight heterogeneous sources, covering 31 body systems, six imaging modalities, and 21 clinical domains across English and seven major Indian languages. We further propose ***ArogyaSutra***, an actor-critic-based multi-agent framework that integrates tool grounding with dual-memory mechanisms for step-wise, reasoning-aware decision making, and uses stored actor-critic simulation trajectories for distillation. Experiments show that our dataset and framework improve multilingual medical reasoning accuracy across all Indic languages, with ablations validating the contribution of each component. The source code and dataset are available at: *https://iitp-cse.github.io/ArogyaSutra/*.

1 Introduction

Multimodal large language models (MLLMs) have demonstrated strong performance on vision–language tasks like multimodal summarization, retrieval, visual question answering, etc and have been adapted for medical applications via supervised fine-tuning on curated multimodal instruction datasets [Li *et al.*, 2023; Jiang *et al.*, 2024b; Acharya *et al.*, 2025]. However, most medical MLLMs rely on a direct-prediction paradigm, producing short answers that fail to capture the interleaved image–text reasoning required in real clinical settings. While reinforcement learning approaches such as GRPO aim to elicit reasoning behaviors, their improvements remain constrained by the underlying base models and do not introduce fundamentally new reasoning paradigms [Guo *et al.*, 2025]. Although chain-of-thought supervision could enhance medical reasoning, curating high-quality medical CoT data is costly and lacks standardized evaluation, motivating the need for methods that can both generate and rigorously evaluate step-by-step multimodal reasoning for complex decision making.

Multimodal Medical reasoning research remains largely English- and Chinese-centric, leaving mid- to low-resource Indic languages severely underrepresented and resulting in uneven reliability across linguistic communities [Qin *et al.*, 2025]. While cross-lingual transfer provides partial relief, open-ended multimodal medical reasoning in Indic languages continues to suffer from two persistent failure modes: *degraded logical fidelity* and *unstable language behavior* [Onyame *et al.*, 2026]. Existing efforts to improve reasoning through domain-specific supervision are further constrained by benchmarks that are largely monolingual, text-centric, and closed-form, providing limited insight into multilingual multimodal reasoning and cultural consistency [NLLB Team, 2024; Surana *et al.*, 2026; Maji *et al.*, 2025]. As multimodal LLMs are increasingly adopted for clinical education and decision support in diverse linguistic settings, systematic evaluation of step-by-step multimodal reasoning and language fidelity in Indic languages is essential for fairness, reliability, and real-world applicability.

To address this gap, we introduce **ArogyaBodha**, a large-scale multimodal medical reasoning dataset and benchmark curated from eight diverse medical sources. **ArogyaBodha** spans seven major Indian languages alongside English and covers multiple imaging modalities, encompassing 31 body systems and 21 clinical domains. Each instance consists of an expert-verified multimodal medical query paired with a single unambiguous ground-truth answer, enabling reliable evaluation of both logical reasoning accuracy and language consistency, as well as systematic analysis of cross-lingual generalization under clinically grounded constraints as shown in Table 1.

To enable reliable multimodal medical reasoning in Indic languages, we propose ***ArogyaSutra***, an actor–critic-based multi-agent framework that combines tool-based visual grounding with dual-memory mechanisms for step-wise, reasoning-aware decision making over image–text inputs. At

each reasoning step t, the **Actor**, implemented as a multimodal language model, processes medical visual inputs and an Indic-language query, invoking lightweight grounding tools [Zhou *et al.*, 2025b] (e.g., zoom/crop, edge detection, depth analysis, and region-level detection) to extract clinically relevant evidence. Rather than directly producing a final answer, the Actor predicts intermediate semantic reasoning steps. The framework is supported by a dual-memory design: *long-term memory* summarizes prior reasoning steps, linguistic formulations, and identified errors (from steps 1 to $t-1$), while *short-term memory* captures the most recent prediction error and feedback. The **Critic** evaluates the Actor's outputs for both medical correctness and language consistency, providing targeted corrective feedback—delivered in English for linguistic errors and in the corresponding Indic language otherwise—which is used to update the memory modules. Through this iterative actor–critic interaction, ***ArogyaSutra*** enables transparent, medically grounded, and linguistically stable reasoning, supporting more trustworthy diagnostic explanations across diverse Indic languages.

To summarize, the contributions of this work can be enumerated as follows: **Problem Formulation.** We formalize a new task of *multimodal medical reasoning in Indic languages*, where a model must perform step-by-step clinical reasoning over medical images and Indic-language queries to produce medically correct and linguistically consistent answers. **Benchmark.** We introduce **ArogyaBodha**, a large-scale multilingual multimodal medical reasoning dataset and benchmark covering seven major Indian languages and English. Curated from eight diverse medical sources, it spans 31 body systems and 21 clinical domains, with expert-verified instances and unambiguous ground-truth answers to enable reliable evaluation of reasoning accuracy. **Framework.** We propose ***ArogyaSutra***, an actor–critic-based multi-agent framework that combines tool-based visual grounding with dual-memory mechanisms for step-wise multimodal medical inference. The framework explicitly tracks and corrects reasoning errors and linguistic inconsistencies through iterative refinement. **Evaluation.** Extensive experiments across languages, imaging modalities, and clinical domains demonstrate that ArogyaSutra consistently outperforms strong multimodal baselines in both reasoning accuracy and multilingual alignment, highlighting its effectiveness for reliable medical AI in Indic language settings.

Impact for AI for Social Good. ***ArogyaSutra*** and **ArogyaBodha** support equitable healthcare access by enabling multilingual multimodal medical reasoning in low-resource Indic languages. The work aligns with AI for Healthcare, Public Health, and Marginalized Communities by promoting clinically grounded and inclusive medical AI systems.

Stakeholders and Multidisciplinary Collaboration. This work is developed in collaboration with certified medical practitioners for clinical validation of the curated medical questions and answers. Translation quality across seven Indic languages is evaluated using reverse-translation analysis with cosine similarity and BLEU_4 metrics to ensure preservation of clinical meaning and linguistic fidelity in multilingual healthcare settings.

Sl.	Dataset	Records	Processing
1	MedXpertQA	1,982	Direct
2	MedTrinity-25M	576	Caption, few-shot
3	MedPix-2.0	464	Report, few-shot
4	MAMA-MIA	64	Caption, few-shot, patient details
5	BRATS24	393	Caption, few-shot, patient details
6	PMC-VQA	1,073	Direct, filtered
7	GMAI-MMBench	164	Direct, filtered
8	NEET-PG, FMGE	391	Extracted, filtered
	Total	**5107**	

Table 1: Summary of datasets of English language after filtering with processing strategies.

2 Related Works

2.1 Medical Multimodal and Multilingual Reasoning

Recent training paradigms such as prompt tuning [Mohanty *et al.*, 2025], supervised fine-tuning (SFT) [Huang *et al.*, 2025], reinforcement learning (RL) [Zhou *et al.*, 2025a] ,agentic distillation [Ghosh *et al.*, 2026] have significantly improved MLLMs on complex vision–language reasoning tasks [Chen *et al.*, 2024] and have been increasingly adapted to medical settings [Huang *et al.*, 2026; Sun *et al.*, 2025]. However, robust multimodal medical reasoning remains challenging due to limitations in supervision strategies and dataset quality [Zhang *et al.*, 2025; Hu *et al.*, 2025]. Prior studies show that SFT with final-answer supervision often leads to overfitting and poor out-of-distribution generalization, as observed in Medvlm-r1 [Pan *et al.*, 2025], while structured reasoning approaches such as Medvlthinker [Huang *et al.*, 2026] remain constrained by modality-specific data. RL-based and multi-agent methods, including Med-R1 [Lai *et al.*, 2025] and MMedAgent-RL [Xia *et al.*, 2026], aim to improve robustness through long-form Chain-of-Thought (CoT) distillation, while Chiron-o1 [Sun *et al.*, 2025] achieves effective reasoning via long-CoT SFT with Mentor–Intern Collaborative Search. Multilingual medical reasoning poses additional challenges, as reasoning performance is strongest in English due to data imbalance [Qiu *et al.*, 2024], and enforcing reasoning in low-resource languages often degrades coherence and amplifies degenerative language patterns [Li *et al.*, 2025].

2.2 Agentic Frameworks in Healthcare

Recent work has explored agentic frameworks to enhance medical reasoning in complex clinical settings. MedAgents[Tang *et al.*, 2024] introduces collaborative multi-agent discussion to improve zero-shot medical reasoning, while MMedAgent[Li *et al.*, 2024] extends this paradigm using reinforcement learning to optimize coordination between specialist agents and a general practitioner agent. Med-R1 [Lai *et al.*, 2025] demonstrates that distilling Chain-of-Thought (CoT) reasoning via reinforcement learning improves robustness when high-quality medical supervision is limited. Complementary efforts such as AgentClinic [Schmidgall *et al.*, 2024] and MedAgentBench [Jiang *et*

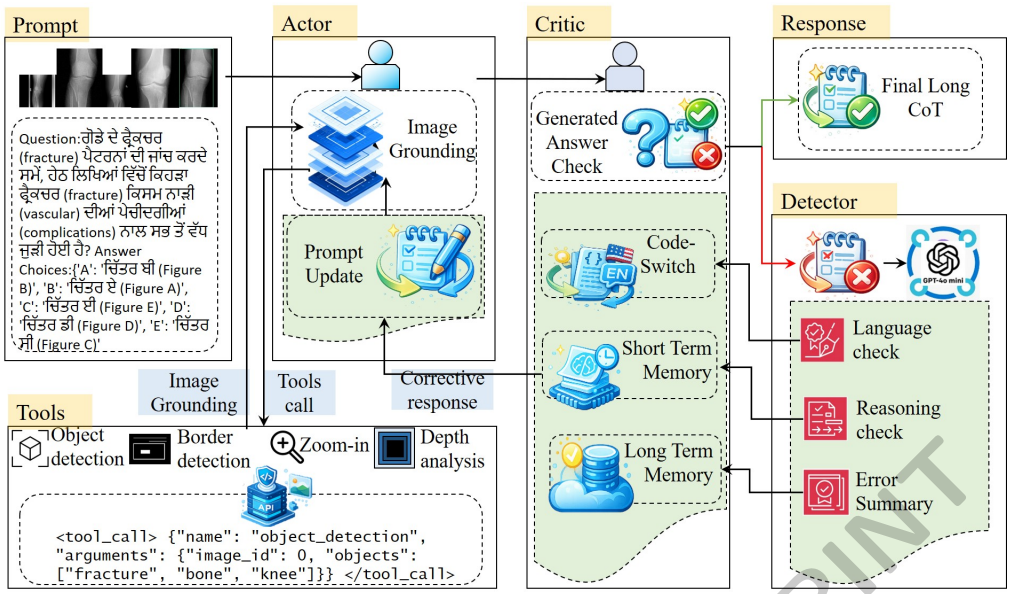

Figure 1: Overview of the ***ArogyaSutra*** framework. ArogyaSutra employs an actor–critic architecture enhanced with tool-based image grounding and adaptive code-switching. The Actor first processes the input prompt and identifies the need for visual grounding, invoking appropriate tool agents to extract clinically relevant information from medical images before generating an answer with its associated reasoning. This output is then passed to the Critic for evaluation. If the response is correct, the Critic approves and outputs the final answer and reasoning. Otherwise, it consults an error detector (GPT-4o-mini) to identify the source of failure. Language-related errors trigger code-switching by translating the query into English, while reasoning-related errors are handled by incorporating summaries of past and current mistakes from long-term and short-term memory. The refined query is then fed back to the Actor for iterative refinement.

al., 2025] focus on realistic evaluation environments for medical agents, emphasizing multi-step decision-making and tool interaction.[Ghosh *et al.*, 2026] improve long-horizon medical reasoning by introducing a dedicated dataset and a distillation-based framework for healthcare task automation. Despite these advances, most existing approaches are limited to high-resource languages or modality-specific settings. Our work builds upon these foundations by introducing an agentic, memory-aware multimodal reasoning framework tailored to low-resource Indic language healthcare scenarios. *To address these gaps—particularly the lack of multimodal reasoning benchmarks for low-resource Indic languages—we curate a dedicated benchmark,* ***ArogyaBodha****. In addition, to enable agentic systems to make informed corrective decisions based on the underlying cause of failure, whether linguistic or logical, we propose a novel Actor–Critic framework,* ***ArogyaSutra****, that adaptively governs feedback and reflection during reasoning.*

3 Dataset: ArogyaBodha

ArogyaBodha comprises 7 major Indian languages namely Bengali(5,111), Hindi(5,104), Assamese(5,108), Tamil(5,106), Telugu(5,106), Punjabi(5,108), and Marathi(5,107), along with English(5,107) curated and filtered from eight heterogeneous medical sources, resulting in a total of 40,857 samples. These sources include benchmark datasets, medical imaging repositories, and postgraduate medical examination questions from India.

3.1 Curation and Filtration

We build the dataset by combining existing medical reasoning benchmarks with questions from Indian postgraduate medical entrance examinations (NEET-PG and FMGE) as shown in Table (Table 1). This ensures that the data is clinically relevant and covers a wide range of medical topics. To better reflect real-world clinical scenarios, we employ few-shot prompting with GPT-4o-mini by combining patient details, reports, and contextual medical information available in the datasets. To keep the questions challenging, The generated questions are further filtered based on reasoning depth and image–question relevance using Gemini-2.5-Flash. Relevance is evaluated using multiple clinical consistency criteria, including modality appropriateness (e.g., CT, MRI, X-ray), anatomical alignment between image and question, visibility of clinical findings, and overall clinical coherence between the image and the associated question. Only samples that pass the relevance checks are included in the final dataset. Finally, the clinical validity and medical correctness of the multimodal questions and answers are reviewed and verified by medical practitioners.

3.2 Translation

The filtered high-quality questions and options are translated into seven major Indian languages using Gemini-2.5-Pro, which is one of the best-known LLMs for generating high-quality translations. This step enables multilingual coverage while preserving the original clinical semantics and reasoning requirements.

Language	BLEU$_4$	Cosine$_{Back}$	Cosine$_{Direct}$
Hindi	0.57	0.94	**0.85**
Marathi	0.50	0.93	0.81
Tamil	0.51	0.93	0.65
Telugu	0.50	0.93	0.64
Bengali	0.51	0.93	0.62
Punjabi	0.56	0.93	0.58
Assamese	0.49	0.93	0.50

Table 2: Reverse (back) translation quality for multilingual medical questions.

3.3 Translation Analysis

Translation quality is evaluated using reverse translation analysis across seven Indic languages, as summarized in Table 2. Semantic consistency between the original English text and translated medical questions is measured using cosine similarity with SentenceTransformer models, while BLEU4 scores are computed between the original and back-translated English text using NLTK. The results show strong semantic preservation across languages, with Cosine$Back$ scores consistently remaining around 0.93–0.94 and higher direct semantic alignment observed for Hindi and Marathi.

To further assess translation quality, 20% of the translated test samples are manually reviewed by medical practitioners proficient in the corresponding regional languages. The evaluators score the translations on a five-point scale based on clinical meaning preservation, medical terminology correctness, semantic consistency, linguistic fluency, and contextual relevance, resulting in an average score of 4.27. This combined automatic and human evaluation framework helps ensure linguistic fidelity, preservation of clinical meaning, and multilingual consistency throughout the benchmark construction process.

4 Framework : *ArogyaSutra*

While recent multimodal models demonstrate strong perceptual grounding in generic environments, their performance degrades substantially in real-world healthcare systems, particularly in low-resource Indic language settings. To address these challenges, ***ArogyaSutra*** introduces a multimodal agentic framework that integrates tool-grounded perceptual reasoning with dual-memory mechanisms comprising short-term and long-term memory to track prior errors and contextual dependencies across reasoning chains. In addition, ***ArogyaSutra*** employs language-aware reflective reasoning to analyze past mistakes and dynamically determine when code-switching is required, enabling robust interaction with complex clinical software in Indic languages. The overall framework is shown in Figure 1

4.1 Vision Tools Grounding

To enhance the perceptual grounding of medical findings, we focus on four visual tools that are particularly effective for perception-centric tasks. Although a broader set of tools is available, empirical observations show that the model consistently underutilizes them during inference. We therefore integrate these four high-performing tools directly into the MLLM inference loop, enabling tool execution and result incorporation to form a coherent multimodal rollout. Specifically, the toolkit includes: (i) open-vocabulary *object detection*, which localizes interface elements given a screenshot and textual query; (ii) *zoom/crop*, which magnifies selected regions for fine-grained inspection; (iii) *Edge Detection*, which extracts the edges of the objects present in the given image; and (iv) *Depth Analysis*, which estimates the relative distances, producing a depth map.

4.2 Memory Utilization

Reasoning errors may arise from linguistic ambiguities or incorrect intermediate logic errors. To prevent the agent from repeating such failures, we introduce an explicit memory mechanism that records past errors and guides future reasoning. The memory is structured into two components: a short-term memory that captures the most recent error, and a long-term memory that summarizes error patterns accumulated over the entire reasoning trajectory. Together, these components enable the agent to account for both immediate mistakes and persistent failure modes when making subsequent decisions. The short-term and long-term memories are represented by the equations below.

$$\mathcal{M}^{(t)}_{\text{short}} = \mathcal{E}_{t-1}, \tag{1}$$

$$\mathcal{M}^{(t)}_{\text{long}} = \text{Summarize}\left(\{\mathcal{E}_0, \ldots, \mathcal{E}_{t-1}\}\right). \tag{2}$$

4.3 Actor-Critic Framework

ArogyaSutra adopts an Actor-Critic formulation as its core decision module, jointly leveraging tool-based perceptual grounding and temporally accumulated memory. Both the Actor and the Critic are instantiated from the same multimodal backbone, Qwen-VL-2.5-7B, and differ only in their functional roles : *action proposal* versus *action evaluation* and their input conditioning. At timestep t, the **Actor** observes the current interface state o_t, task objective g, perceptual context z_t, and memory states $(\mathcal{M}^S_t, \mathcal{M}^L_t)$, and samples a candidate semantic action:

$$u_t \sim \pi_\theta\big(u_t \mid o_t, g, z_t, \mathcal{M}^S_t, \mathcal{M}^L_t\big), \tag{3}$$

where π_θ denotes the Actor policy and u_t represents the proposed action. The *Critic*, parameterized by ψ, evaluates the proposed action by estimating a scalar confidence score:

$$V_\psi\big(o_t, g, u_t, \mathcal{M}^S_t, \mathcal{M}^L_t\big) \to \hat{s}_t \in [0, 1], \tag{4}$$

where $\hat{s}_t$ reflects the estimated correctness of the action. If $\hat{s}_t = 1$, the action is accepted, and both memory modules are updated; otherwise, the Critic triggers *language-Analysis reflection* and emits structured corrective feedback Δ_t to guide the subsequent reasoning chain.

Language Analysis Aware Reflection. When the Actor produces an incorrect prediction, the Critic first diagnoses the source of the failure, distinguishing between language understanding errors, such as repeated tokens or incomplete generations, and logical errors in the reasoning chain. If the error is attributed to language instability, the Critic provides reflective feedback in English to stabilize and guide the Actor's

Model	Assamese	Bengali	Hindi	Marathi	Punjabi	Tamil	Telugu	Average
GPT 4.0 mini	35.38	45.38	37.69	33.07	36.15	31.53	33.20	36.05
GPT 4.0	38.40	44.60	40.10	37.80	39.20	36.90	39.10	39.30
Qwen3-VL-8B-Instruct	41.30	40.50	41.50	41.50	41.60	38.70	39.90	40.71
Mistral-Small-3.2-24B-Instruct	40.90	42.90	44.10	43.20	40.90	41.40	42.50	42.27
LLaVA-v1.6-34B	32.09	31.45	34.44	33.66	31.70	31.31	35.42	32.87
BioMistral-7B	25.64	25.78	28.38	25.83	24.27	27.59	26.95	26.35
Llama3-OpenBioLLM-8B	16.10	15.60	22.20	20.50	11.70	12.10	7.70	15.13
MedGemma-4B-it	38.50	35.10	37.40	34.50	35.70	36.70	34.90	36.11
MedVLM-R1	23.50	22.80	25.60	22.80	24.20	25.40	22.20	23.79
Qwen2.5-VL-3B-Instruct	28.70	30.70	31.50	31.30	28.40	28.20	28.10	29.56
Qwen2.5-VL-7B-Instruct	32.30	36.50	33.84	36.12	32.30	35.38	33.07	34.21
ArogyaSutra(Qwen 2.5 VL 3B)	34.69	37.08	38.36	37.03	33.07	35.73	33.60	35.65
ArogyaSutra(Qwen 2.5 VL 7B)	47.69	45.38	42.30	42.30	43.07	41.53	41.53	43.40

Table 3: Comparison of ***ArogyaSutra*** with respect to baselines on various Indic languages. The best results per column are highlighted in underline. The last group reports the overall Average. * denotes our proposed method.

Model	Assamese	Bengali	Hindi	Marathi	Punjabi	Tamil	Telugu	Average
ArogyaSutra*(with all components)*	47.69	45.38	42.30	42.30	43.07	41.53	41.53	43.40
w/o Critic and Image Grounding	40.00	36.00	34.00	32.00	32.00	28.00	32.00	33.43
w/o Critic	42.00	32.0	36.00	30.00	34.00	30.00	32.00	33.71
w/o Image Grounding	24.00	28.00	32.00	28.00	26.00	22.00	28.00	26.86
w/o code-switch	30.00	32.00	26.00	24.00	28.00	24.00	24.00	26.86

Table 4: Results on various ablations in ***ArogyaSutra*** show the impact of Image Grounding, critic agent, and code-switching.

reasoning chain to rectify itself. Conversely, if the failure stems from faulty reasoning, the Critic issues the reflection in the corresponding Indic language of the query, enabling more contextually grounded corrective guidance.

Halting Condition. If the actor–critic framework converges to the correct answer within three or fewer reasoning iterations, the resulting reasoning trace is retained. Otherwise, the reasoning process is reset and restarted using a summary of the accumulated errors stored in long-term memory. If the framework fails to reach the correct answer in two consecutive restart cycles, the corresponding data point is deemed unreliable and is removed from the training set.

4.4 Training Strategy

To progressively enhance the Actor's capabilities, we perform code-switched reasoning chain distillation by transferring the Critic's corrective signals into the Actor, thereby removing the need for explicit evaluation at inference time. Specifically, after simulating Actor-Critic rollouts, we construct a Critic-refined dataset $\mathcal{D} = \{(o_i, g_i, z_i, \mathcal{M}_i^S, \mathcal{M}_i^L, u_i^\dagger)\}_{i=1}^N$, where $u_i^\dagger$ denotes the Critic-approved action augmented with relevant tool outputs and memory context for the subsequent step.

The Actor is then fine-tuned using supervised learning with the following objective:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{N}\sum_{i=1}^{N} \log \pi_\theta\Big(u_i^\dagger \mid o_i, g_i, z_i, \mathcal{M}_i^S, \mathcal{M}_i^L\Big). \quad (5)$$

During inference, only the distilled policy π_θ is retained. Conditioned on the current interface state, task objective, and memory signals, the Actor directly predicts the next semantic action without invoking the Critic. This distillation strategy preserves the Critic's structured reasoning and memory-aware behavior while substantially reducing inference-time overhead.

5 Experimental Setup

This section describes our experimental setups. First section describes implementation details, second section describes evaluation metrics and the last section is for baselines.

5.1 Implementation Details

All experiments were conducted on an NVIDIA A100 80GB PCIe GPU. Models were trained for 3 epochs on an 18k multilingual multimodal dataset with balanced language proportions, requiring approximately 12–15 hours per run. We used PyTorch, Hugging Face Transformers, and Unsloth for efficient training. The learning rate was set to 5×10^{-4} with 100 warmup steps and a weight decay of 0.01. Training was performed with a per-device batch size of 2 and 4 gradient accumulation steps. We fine-tuned the Qwen2.5-VL-7B-Instruct model using Unsloth with gradient checkpointing and 4-bit precision (`load_in_4bit=True`) for memory efficiency. Parameter-efficient fine-tuning was applied to the language, attention, and MLP projection layers, with LoRA rank $r = 16$, LoRA $\alpha = 16$, and dropout set to 0.

5.2 Testset Composition

From the curated dataset, we select 130 samples per language to construct the test set, resulting in 910 samples across seven

Model	Accuracy
Qwen3-VL-8B-Instruct	49.6
Qwen2.5-VL-3B-Instruct	38.6
LLaVA-v1.6-34B	23.3
BioMistral-7B	36.9
Llama3-OpenBioLLM-8B	39.3
MedGemma-4B-it	45.2
MedVLM-R1	23.6
Qwen2.5-VL-7B-Instruct (Baseline)	35.0
ArogyaSutra(Qwen 2.5 VL 7B)	**50.4**

Table 5: Performance of ***ArogyaSutra*** on OOD dataset.

languages. The test set is balanced across data sources, with identical questions used for all languages to enable controlled cross-lingual comparison. Each language's test set comprises 42 samples from MedXpertQA, 34 from PMC-VQA, 16 from MedPix-2.0, 17 from MedTrinity-25M, 9 from BRATS24, 8 from NEET-PG and FMGE, 3 from GMAI-MMBench, and 1 from MAMA-MIA. We curate a set of 60 multimodal medical queries derived from real-world clinical examinations of the Médico Interno Residente (MIR) residency exams in Spain [Riccio *et al.*, 2025], each translated into all seven evaluated Indic languages to enable controlled cross-lingual out-of-distribution (OOD) evaluation.

5.3 Baselines and Evaluation Metric

We evaluate the proposed question set across a diverse range of medical and non-medical multimodal large language models (MLLMs) to ensure a comprehensive comparison. Our evaluation includes vision-language models from the Qwen family[Wang *et al.*, 2024] (Qwen3-VL-8B-Instruct and Qwen2.5-VL-3B-Instruct) to analyze performance across model scales, as well as strong general-purpose MLLMs—Mistral-Small-3.2-24B-Instruct[Jiang *et al.*, 2024a], and LLaVA-v1.6-34B[Liu *et al.*, 2023] which are not specifically trained for medical reasoning. For domain-focused assessment, we additionally, benchmark medical MLLMs, including BioMistral-7B[Labrak *et al.*, 2024], Llama3-OpenBioLLM-8B[Dorfner *et al.*, 2024], MedGemma-4B-it[Sellergren *et al.*, 2025], and MedVLM-R1[Pan *et al.*, 2025]. Among closed source models, we use GPT 4.0 mini and GPT 4.0. All models are evaluated in a zero-shot setting, and performance is measured using choice accuracy as the primary metric.

6 Result and Analysis

We evaluate ***ArogyaSutra*** against strong *general-purpose* and *medical-domain* multimodal baselines. As shown in Table 3, integrating tool grounding with short- and long-term memory substantially improves multimodal medical reasoning. We further conduct controlled component and robustness analyses on an out-of-distribution benchmark [Riccio *et al.*, 2025] medical questions across multiple Indic languages.

6.1 Research Question

a) How does *ArogyaSutra* perform compared to baselines? Table 3 shows that ***ArogyaSutra*** achieves the best average performance across evaluated languages. The Qwen2.5-VL-7B variant achieves the highest average accuracy of **43.40**, surpassing the strongest baseline (GPT-4.0 at 39.30) by **+4.1 points**, with consistent gains across all seven languages. It also significantly outperforms medical-specific models such as *BioMistral-7B* (23.83 avg) and *MedGemma-4B-it* (36.11 avg), highlighting a clear advantage over domain-specialized baselines. Notably, ***ArogyaSutra*** improves over its base Qwen2.5-VL-7B model by **+9.2 points**, demonstrating the effectiveness of the proposed approach for robust cross-lingual medical reasoning.

b) How is the performance of *ArogyaSutra* compared to the baseline models it has been trained on? In Table 3, ***ArogyaSutra*** substantially outperforms the baseline models on which it is trained. ***ArogyaSutra*** (Qwen2.5-VL-3B) improves the average accuracy from 29.56 to 35.65, yielding a gain of **+6.1 points** over Qwen2.5-VL-3B-Instruct. More prominently, ArogyaSutra (Qwen2.5-VL-7B) achieves an average accuracy of **43.40**, surpassing its base Qwen2.5-VL-7B-Instruct model (34.21) by **+9.2 points**, with consistent improvements across all evaluated languages. This indicates that ***ArogyaSutra*** delivers substantial gains beyond its underlying baseline models.

c) How is the performance of *ArogyaSutra* on out of distribution dataset? Table 5 shows the performance of ***ArogyaSutra*** on OOD dataset. On the out-of-distribution (OOD) dataset [Riccio *et al.*, 2025], ***ArogyaSutra*** demonstrates strong generalization performance. ***ArogyaSutra***(Qwen2.5-VL-7B) achieves an accuracy of **50.4**, outperforming its base model Qwen2.5-VL-7B-Instruct baseline (35.0) by a clear margin. It also surpasses large general-purpose MLLMs such as LLaVA-v1.6-34B (23.3), as well as medical-specific models including BioMistral-7B (36.9) and MedGemma-4B-it (45.2). These results indicate that ***ArogyaSutra*** exhibits robust reasoning and improved resilience under distribution shift.

7 Ablation Studies

Impact of the Critic Agent and its Components. Removing the Critic significantly reduces performance, with Actor-only variants achieving 33.43%–33.71% compared to **43.40%** for the full *ArogyaSutra* framework (Table 4). Even the grounded Actor-only model underperforms the baseline *Qwen2.5-VL-7B-Instruct* (34.21%). Furthermore, removing tool grounding or code-switching causes the largest performance drop (16.54%), while disabling long- and short-term memory reduces accuracy by 12.83%. These results highlight the importance of reflective error correction, grounding, memory, and adaptive multilingual reasoning in *ArogyaSutra*.

Impact of Tool Grounding and Memory. Table 6 shows that disabling both components reduces accuracy to 26.86%. Memory alone improves performance to 30.57%, while enabling both achieves the best accuracy of **43.40%**, demonstrating their complementary role in multilingual multimodal reasoning.

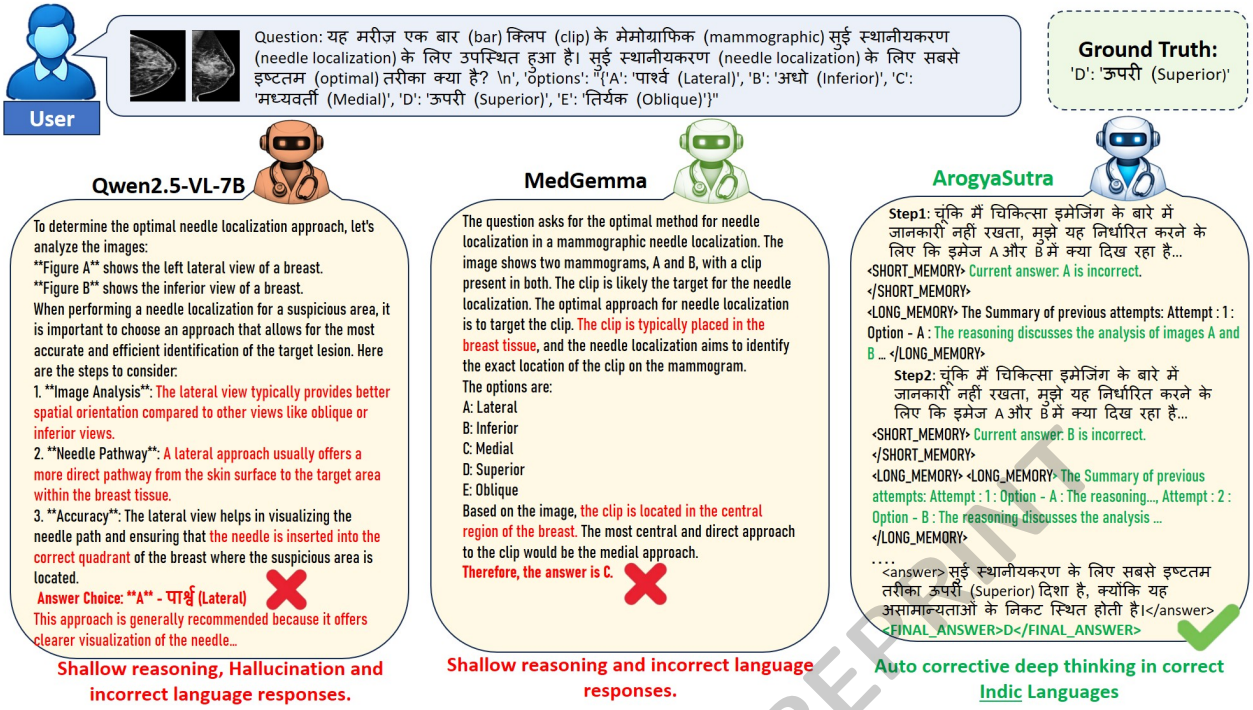

Figure 2: Overview of qualitative comparisons between ***ArogyaSutra*** and baseline models.

Image Grounding	Memory Utilization	Average Output (%)
✗	✓	26.86
✓	✗	30.57
✓	✓	**43.40**

Table 6: Ablation study of image grounding and memory utilization within the Critic agent. Best performance is achieved when both components are enabled.

8 Qualitative Analysis

We qualitatively analyze the generated outputs of the proposed model and compare them with two representative baselines: the medical MLLM *MedGemma-4B-it* and the general-purpose vision–language model *Qwen2.5-VL-7B-Instruct* as shown in the Figure 2. Both baselines exhibit shallow reasoning and frequently fail to respond in the input low-resource language (e.g., Hindi). In contrast, ***ArogyaSutra*** consistently generates responses in the target language and demonstrates more structured and corrective reasoning. Leveraging the actor-critic framework, the model identifies and revises incorrect intermediate conclusions, progressively eliminating implausible hypotheses. These observations highlight the effectiveness of ***ArogyaSutra*** in improving multilingual alignment and reasoning fidelity in multilingual medical settings.

9 Risk Analysis and Limitations

Despite the strong performance of ***ArogyaSutra*** and the **ArogyaBodha** benchmark, several limitations remain. The model may still produce reasoning errors or misinterpret visual evidence, particularly in rare or atypical clinical cases, which poses risks if outputs are misused for real-world decision making. While **ArogyaBodha** covers seven major Indian languages, it does not capture the full linguistic diversity of India, including low-resource dialects and code-mixed language commonly used in clinical practice, potentially degrading performance under such conditions. Finally, the actor–critic framework depends on the reliability of underlying visual grounding tools and the base multimodal backbone; failures in perception or error attribution can propagate through the reasoning process despite critical feedback.

10 Conclusion

We introduce **ArogyaBodha** and ***ArogyaSutra*** to advance multilingual multimodal medical reasoning in low-resource Indic-language settings, demonstrating consistent improvements over strong general-purpose and medical baselines across languages and imaging modalities. By integrating tool-grounded perception, actor–critic reasoning, memory-aware reflection, and adaptive code-switching, ***ArogyaSutra*** enables more reliable and linguistically consistent medical reasoning across diverse Indic languages. Extensive experiments, ablation studies, and out-of-distribution evaluations demonstrate the effectiveness and robustness of the proposed framework. We hope this work encourages future research in trustworthy multilingual medical AI and clinically grounded multimodal reasoning systems for underserved communities.

Ethical Statement

All datasets are anonymized and contain no personally identifiable patient information. All datasets used in **Arogya-Bodha**, including MedXpertQA, MedTrinity-25M, PMC-VQA, MedPix-2.0, MAMA-MIA, BraTS 2024, GMAI-MMBench, and NEET-PG/FMGE, are publicly available under their respective research, academic, or open-access licenses and are utilized solely for non-commercial research purposes.

Acknowledgments

The authors gratefully acknowledge the Aryabhatta Supercomputing Centre (ASC) at the Indian Institute of Technology Patna, established under the National Supercomputing Mission (NSM), Government of India, for providing the computational resources utilized in this work. The authors also sincerely thank Dr. Muhsin Muhsin from IGIMS and Dr. Priti Singh from IIT Patna for their valuable collaboration and support throughout the project, particularly in the data curation and verification processes.

Contribution Statement

Tanmoy Kanti Halder and Akash Ghosh contributed equally to this work.

References

[Acharya *et al.*, 2025] Arkadeep Acharya, Akash Ghosh, Pradeepika Verma, Kitsuchart Pasupa, Sriparna Saha, and Priti Singh. M3retrieve: Benchmarking multimodal retrieval for medicine. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15274–15287, 2025.

[Chen *et al.*, 2024] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198. IEEE/CVF, 2024.

[Dorfner *et al.*, 2024] Felix J. Dorfner, Amin Dada, Felix Busch, Marcus R. Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C. Adams, and Keno K. Bressem. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *CoRR*, abs/2408.13833, 2024.

[Ghosh *et al.*, 2026] Akash Ghosh, Tajamul Ashraf, Rishu Kumar Singh, Numan Saeed, Sriparna Saha, Xiuying Chen, and Salman Khan. Carepilot: A multi-agent framework for long-horizon computer task automation in healthcare. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9695–9705, 2026.

[Guo *et al.*, 2025] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[Hu *et al.*, 2025] Yuan Hu, Chenhan Xu, Bo Lin, Weibin Yang, and Yuan Yan Tang. Medical multimodal large language models: A systematic review. *Intell. Oncol.*, 1(4):308–325, October 2025.

[Huang *et al.*, 2025] Chao Huang, Zeliang Zhang, Jiang Liu, Ximeng Sun, Jialian Wu, Xiaodong Yu, Ze Wang, Chenliang Xu, Emad Barsoum, and Zicheng Liu. Directional reasoning injection for fine-tuning MLLMs, 2025.

[Huang *et al.*, 2026] Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. MedVLThinker: Simple baselines for multimodal medical reasoning. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pages 384–398. PMLR, 13–14 Dec 2026.

[Jiang *et al.*, 2024a] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

[Jiang *et al.*, 2024b] Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. Med-MoE: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3843–3860, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Jiang *et al.*, 2025] Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *Nejm Ai*, 2(9):AIdbp2500144, 2025.

[Labrak *et al.*, 2024] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Lai *et al.*, 2025] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-R1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE Journals & Magazine*, 2025.

[Li *et al.*, 2023] 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.

[Li *et al.*, 2024] Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. MMedAgent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

[Li *et al.*, 2025] Feiyang Li, Yingjian Chen, Haoran Liu, Rui Yang, Han Yuan, Yuang Jiang, Tianxiao Li, Edison Marrese Taylor, Hossein Rouhizadeh, Yusuke Iwasawa, Douglas Teodoro, Yutaka Matsuo, and Irene Li. MKG-Rank: Enhancing large language models with knowledge graph for multilingual medical question answering, 2025.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[Maji *et al.*, 2025] Arijit Maji, Raghvendra Kumar, Akash Ghosh, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, Sriparna Saha, et al. Drishtikon: A multimodal multilingual benchmark for testing language models' understanding on indian culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1289–1313, 2025.

[Mohanty *et al.*, 2025] Anwesha Mohanty, Venkatesh Balavadhani Parthasarathy, and Arsalan Shahid. The future of MLLM prompting is adaptive: A comprehensive experimental evaluation of prompt engineering methods for robust multimodal performance, 2025.

[NLLB Team, 2024] NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024.

[Onyame *et al.*, 2026] Eric Onyame, Akash Ghosh, Subhadip Baidya, Sriparna Saha, Xiuying Chen, and Chirag Agarwal. Cure-med: Curriculum-informed reinforcement learning for multilingual medical reasoning, 2026.

[Pan *et al.*, 2025] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2025*. Springer, 2025.

[Qin *et al.*, 2025] Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. A survey of multilingual large language models. *Patterns*, 6(1), 2025.

[Qiu *et al.*, 2024] Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards building multilingual language model for medicine. *Nat. Commun.*, 15(1):8384, September 2024.

[Riccio *et al.*, 2025] Giuseppe Riccio, Antonio Romano, Mariano Barone, Gian Marco Orlando, Diego Russo, Marco Postiglione, Valerio La Gatta, and Vincenzo Moscato. A multilingual multimodal medical examination dataset for visual question answering in healthcare. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 435–440, 2025.

[Schmidgall *et al.*, 2024] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments, 2024.

[Sellergren *et al.*, 2025] 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, 2025.

[Sun *et al.*, 2025] Haoran Sun, Yankai Jiang, Wenjie Lou, Yujie Zhang, Wenjie Li, Lilong Wang, Mianxin Liu, Lei Liu, and Xiaosong Wang. Chiron-o1: Igniting multimodal large language models towards generalizable medical reasoning via mentor-intern collaborative search. In *Advances in Neural Information Processing Systems*, 2025.

[Surana *et al.*, 2026] Harshul Raj Surana, Arijit Maji, Aryan Vats, Akash Ghosh, Sriparna Saha, and Amit Sheth. Viraasat: Traversing novel paths for indian cultural reasoning, 2026.

[Tang *et al.*, 2024] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, 2024.

[Wang *et al.*, 2024] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

[Xia *et al.*, 2026] Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Shujie Liu, Yan Lu, and Huaxiu Yao. MMedAgent-RL: Optimizing multi-agent collaboration for multimodal medical reasoning. In *International Conference on Learning Representations*, 2026.

[Zhang *et al.*, 2025] Andrew Zhang, Eric Zhao, Ruirui Wang, Xiuqi Zhang, Justin Wang, and Ethan Chen. Multimodal large language models for medical image diagnosis: Challenges and opportunities. *J. Biomed. Inform.*, 169(104895):104895, September 2025.

[Zhou *et al.*, 2025a] Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced MLLM: A survey on RL-based reasoning in multimodal large language models, 2025.

[Zhou *et al.*, 2025b] Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu, Mingyang Fu, Sinan Wang, Yao Wan, Zhou Zhao, and Ranjay Krishna. Reinforced visual perception with tools, 2025.