

ARISE: Adaptive Agentic Reasoning with Image-grounded Self-Evaluation for Interpretable IBD Assessment

Pronoma Banerjee¹, Anuva Shah¹, Jason Wu¹, Md. Masudur Rahman¹, Sanjay Mohanty², Satya Kurada², and Juan P. Wachs¹

¹ Purdue University, West Lafayette, IN 47907, USA

² Indiana University School of Medicine, Indianapolis, IN 46202, USA
jpwachs@purdue.edu

Abstract

Inflammatory bowel disease (IBD) requires frequent imaging-based assessment, yet interpretation of modalities such as wireless capsule endoscopy (WCE) and intestinal ultrasound remains heavily dependent on specialist expertise. Vision-Language Models (VLMs) demonstrate significant potential in multimodal medical image analysis, but their clinical adoption is hindered by their insufficient domain-specific reasoning, susceptibility to hallucination, scarcity of good quality training data in fine-grained diagnostics and limited interpretability. We introduce ARISE (Adaptive Agentic Reasoning with Image-grounded Self-Evaluation), an autonomous planning framework that models few-shot medical image understanding as a sequential agentic workflow. ARISE structures agent execution into a transparent 5-stage workflow: hypothesis generation, image-grounded evidence summarization, evidence-conditioned refinement, symbolic verification, and final diagnosis. We apply ARISE to IBD assessment across two independent patient cohorts: wireless capsule endoscopy (WCE) images for Crohn’s disease and B-Mode ultrasound data for Ulcerative Colitis. ARISE consistently improves diagnostic performance over baseline VLMs while exposing where reasoning succeeds or fails, providing a more interpretable basis for clinical decision support and realistic deployment.

Introduction

Inflammatory bowel disease (IBD), including Crohn’s disease (CD) and ulcerative colitis (UC), is a chronic relapsing condition that requires repeated assessment of disease activity to prevent progressive bowel damage and other complications (Ng et al. 2017; Turner et al. 2021; Torres et al. 2017). Two imaging modalities are central to this monitoring workflow. Wireless capsule endoscopy (WCE) (Iddan et al. 2000) provides direct visualization of small-bowel mucosal abnormalities in Crohn’s disease, while intestinal ultrasound (IUS) (Maaser et al. 2020) offers a radiation-free means of assessing bowel-wall features such as colon wall thickening and loss of stratification, both established markers of activity in ulcerative colitis (Bots et al. 2018). Despite their clinical value, both modalities share a critical bottleneck: accurate interpretation depends heavily on scarce, subspecialty-trained experts (Vallée et al. 2020; Leenings et al. 2025). In resource-limited and non-tertiary settings, this dependence

delays diagnosis and monitoring precisely where timely intervention matters most (Kaplan and Windsor 2021).

Vision-language models (VLMs) offer a promising way to support medical image interpretation without task-specific training (Radford et al. 2021; Alayrac et al. 2022; Liu et al. 2023; Achiam et al. 2023). Yet their translation into IBD assessment is hindered by four well-documented obstacles (Hager et al. 2024). First, general-purpose VLMs exhibit a domain-reasoning gap: fine-grained IBD diagnosis requires integrating subtle, co-occurring visual cues over disease progression rather than recognizing salient objects, and these cues are underrepresented in general training distributions (Du, Yu, and Wang 2025; Zhang et al. 2025). Second, VLMs remain prone to hallucination and to internally contradictory rationales, an unacceptable failure mode in a clinical setting (Ji et al. 2023; Li et al. 2025; Madaan et al. 2023). Third, high-quality annotated IBD imaging data are scarce, ruling out conventional fine-tuning in most practical deployments (Shakeri et al. 2024; Vallée et al. 2020). Finally, even accurate VLMs act as black boxes, providing no way to distinguish sound visual reasoning from chance or spurious shortcut (Byun et al. 2025, 2026); inference-time methods such as test-time prompt tuning and retrieval-augmented in-context learning avoid retraining, but do not necessarily produce reasoning that can be inspected or verified (Shu et al. 2022; Ram et al. 2023; Zhan et al. 2026; Xia et al. 2025).

Neurosymbolic methods that translate natural-language problems into symbolic form for solver-backed verification improve logical faithfulness, but they have been developed for text-only reasoning rather than image-grounded medical diagnosis, where hypotheses must be tied to observable visual findings (Pan et al. 2023; Olausson et al. 2023; Garcez and Lamb 2023). Moreover, most medical-agent evaluations emphasize final-task accuracy, making it difficult to identify where failures arise within a multi-stage reasoning process (Liu et al. 2024; Fu et al. 2025; Hager et al. 2024). As recent workflow-aware benchmarks have argued, final-output metrics alone cannot reveal where or why an agent fails (Schmidgall et al. 2024; Jiang et al. 2025). In short, no existing few-shot adaptation approach produces a reasoning state that is simultaneously grounded in the image evidence, verified for internal consistency, and open to expert inspection before deployment.

We introduce ARISE (Adaptive Agentic Reasoning with

Image-grounded Self-Evaluation), an autonomous framework that adapts a frozen VLM to few-shot IBD assessment by constructing, grounding, verifying, and refining an explicit set of diagnostic hypotheses, and that evaluates both its intermediate reasoning stages and its final diagnosis. ARISE follows five stages: (S1) generation of diagnostic hypotheses from task and clinical knowledge, (S2) image-grounded summarization of findings from a small support set, (S3) evidence-conditioned refinement of those hypotheses, (S4) symbolic verification and repair using first-order logic and an SMT solver, and (S5) final diagnosis of unseen images. This design makes transparency and correctness mutually reinforcing, and constrains hallucination without fine-tuning. Each stage produces an inspectable artifact and an associated score, allowing ARISE to evaluate both the quality of its own reasoning process and the accuracy of the resulting diagnosis evaluated in two settings: Wireless Capsule Endoscopy (WCE) for Crohn’s disease and B-mode ultrasound for ulcerative colitis.

Our contributions are fourfold: First, we introduce a few-shot adaptation strategy in which domain knowledge is represented as image-grounded, iteratively refined hypotheses rather than learned parameters. Second, we extend solver-backed neurosymbolic reasoning to image-grounded medical diagnosis by verifying and repairing VLM-generated hypotheses before prediction. Third, we introduce stage-level evaluation that separates agentic process quality from task performance, revealing where reasoning failures occur rather than reducing evaluation to a final label. Finally, we demonstrate the framework across independent WCE (Vallée et al. 2020) and B-mode ultrasound (Leenings et al. 2025) cohorts and multiple VLM families, showing that structured, verifiable reasoning can improve diagnostic performance across modalities and model backbones.

Methodology

We describe **ARISE** (*Adaptive Agentic Reasoning with Image-grounded Self-Evaluation*), an autonomous agent that adapts a frozen vision-language model (VLM) to few-shot inflammatory bowel disease (IBD) assessment by executing a transparent, five-stage reasoning workflow. Unlike a single opaque prompt, ARISE carries an explicit workflow *state* across stages, invokes an external symbolic tool, and emits an inspectable artifact at every stage that receives its own *score*. No VLM parameters are updated: adaptation is stored entirely as an auditable hypothesis state. This eliminates the dependency on massive datasets, bypassing the conventional fine-tuning bottleneck that typically hinders medical AI deployment. We summarize a run as an *agentic* (process) score over the five workflow stages plus a *task* (outcome) score, and combine them into a single overall score.

Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be an IBD imaging dataset, where x_i is an input image (a wireless capsule endoscopy (WCE) frame for the Crohn’s cohort or a B-mode ultrasound frame for the ulcerative colitis (UC) cohort) and $y_i \in \mathcal{Y}$ is a diagnostic label. We operate in the few-shot regime: a small

labeled support set $\mathcal{S} = \{(x_j, y_j)\}_{j=1}^m$ is available for adaptation, and a disjoint query set $\mathcal{Q}$ is held out. Each task carries a task description T (modality, label space $\mathcal{Y}$, acquisition context, requested output) and a clinical cue description C enumerating expert knowledge about the visual findings associated with each label.

The **ARISE agent** is the tuple $(M_\theta, \pi, V, \Gamma)$: M_θ is the frozen VLM; π is a deterministic controller that sequences the stages and triggers repair loops; V is a symbolic satisfiability-modulo-theories (SMT) verifier (Z3); and Γ is a prompt-based parser that maps each natural-language hypothesis to a typed first-order-logic (FOL) premise. The agent maintains a state

$$\sigma^{(t)} = (H^{(t)}, \Phi^{(t)}, \mathcal{E}^{(t)}, \mathbf{z}^{(t)}), \quad (1)$$

comprising the hypothesis set $H^{(t)}$, its premises $\Phi^{(t)} = \Gamma(H^{(t)})$, an evidence memory $\mathcal{E}^{(t)}$, and the vector of stage scores $\mathbf{z}^{(t)} = (s_1, \dots, s_5)$. The exact scores are described in details in Appendix A. The goal is to construct a verified hypothesis state H^* that yields an accurate query prediction $\hat{y} = f_{\text{task}}(x, H^*, \mathcal{E}^*, M_\theta)$ while producing an inspectable, individually-scored reasoning trace.

Stage S1: Hypothesis Generation

ARISE composes a planning prompt from task and clinical knowledge, $p_0 = \text{Plan}(T, C, \mathcal{Y})$, and the VLM generates concise, checkable hypotheses

$$H^{(0)} = \{h_1, \dots, h_K\} = M_\theta(p_0; \tau, p_{\text{top}}), \quad (2)$$

where the sampling temperature τ and nucleus sampling parameter p_{top} can be varied to encourage hypothesis diversity. For example, in the WCE six-lesion task, p_0 may ask: “Given Crohn’s disease capsule endoscopy images and the label set $\{\text{AU}, \text{E}, \text{O}, \text{S}, \text{U3} - 10, \text{U} > 10\}$, generate concise visual hypotheses linking observable mucosal findings to each label. Consider mucosal breaks, white ulcer base, red halo, diffuse redness, enlarged villi, and luminal narrowing. Avoid claims that cannot be checked from the image.”

Stage S2: Image-Grounded Evidence Summarization

ARISE initializes an empty evidence memory $\mathcal{E}^{(0)} = \emptyset$ and picks one example image at a time, sequentially. At step t it selects (x_t, y_t) and asks the VLM to report only the *observable* evidence in x_t supporting y_t under the current hypotheses:

$$e_t = M_\theta(\text{Observe}(x_t, y_t, p_0)). \quad (3)$$

The output is *structured*: e_t tags each cited finding against a per-cohort controlled vocabulary $\mathcal{V}$ (for endoscopy, this will be mucosal break, white ulcer base, red halo, diffuse erythema, enlarged villi, luminal narrowing, etc.). This makes scoring a set comparison rather than text parsing, which is simpler to evaluate. The generated visual summary is appended as $\mathcal{E}^{(t)} = \mathcal{E}^{(t-1)} \cup \{e_t\}$. The final evidence set is denoted as $\mathcal{E}^*$. The grounding score s_2 measures agreement between model-identified findings and expert-annotated findings, emphasizing visual fidelity over unsupported inference.

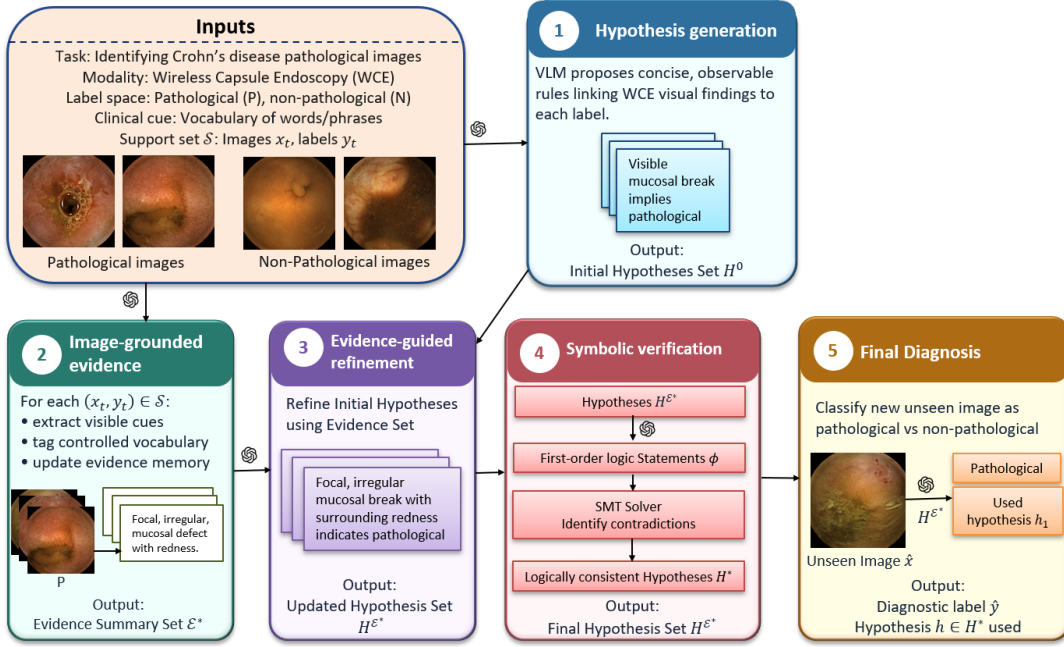


Figure 1: An overview of the ARISE workflow for the Crohn's disease WCE Pathological vs Non-Pathological image classification. The workflow starts with initial hypotheses generation and a parallel summarization of visual cues from select examples. The initial hypotheses are first updated using the summary of visual cues, followed by a logical consistency check via an SMT solver. The final hypotheses are used for the final diagnosis stage.

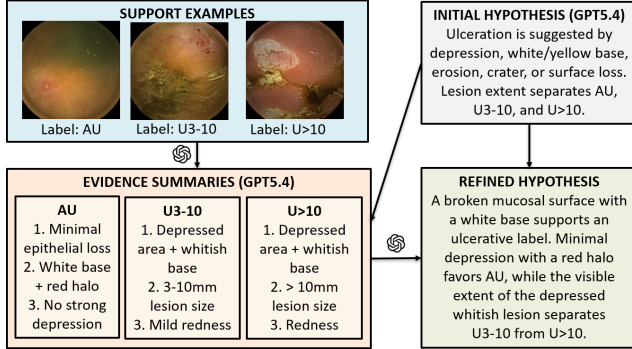


Figure 2: Example of stages S2 and S3 on the Crohn's disease Ulceration Diagnosis task with GPT-5.4

Stage S3: Evidence-Conditioned Refinement

Given the accumulated evidence, the agent proposes a refined set $H^{(\mathcal{E}^*)} = M_\theta(\text{Refine}(p_0, H^{(0)}, \mathcal{E}^*))$. The refinement score s_3 measures whether the refined hypothesis set improves downstream prediction performance relative to the initial hypotheses on a held-out validation subset.

Stages S2 and S3 are jointly described in Figure 2.

Stage S4: Symbolic Verification

The generated hypotheses $H^{(\mathcal{E}^*)}$ are converted from natural language to first order logic statements Φ using an LLM parser Γ , $\Phi = \Gamma(H^{(\mathcal{E}^*)})$. For example, the nat-

ural language statement "A mucosal break with a white base indicates aphthoid ulceration" is translated to the form $(\text{MucosalBreak}(x) \wedge \text{WhiteBase}(x)) \rightarrow \text{AU}$. These are then passed into a Satisfiability Modulo Theories Solver.

$$V(\Phi) = \begin{cases} 1, & \text{if } \Phi \text{ is satisfiable,} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

If $V(\Phi) = 0$, V returns the minimal conflicting set b and the LLM repairs the hypotheses, $H^* = M_\theta(\text{Repair}(p_0, H^{(\mathcal{E}^*)}, b))$, giving the final hypotheses set H^* to be used for diagnosis.

The verification score s_4 measures the fidelity of the natural-language-to-logic translation, evaluated through bidirectional entailment between the original and reconstructed hypotheses. This is done using a natural language inference model, details are provided in the Appendix A.

Stage S5: Final Diagnosis

With the verified state H^* , the model performs a downstream classification task, classifying an input $\hat{x}$ with a corresponding label $\hat{y}$. A few examples of pictures and labels are added for in-context learning. For each classification, we also output the hypothesis h it uses for the classification. The stage score s_5 measures whether the prediction belongs to the valid label space and whether the model grounds its prediction in the verified hypothesis set.

The task score for the entire agentic system is measured as the average of the classification accuracy, F1-score, precision, and recall on $\mathcal{Q}$.

Workflow-Aware Scoring: Agentic + Task

The five workflow stages are aggregated into a single *agentic* (process) score, and the outcome is the *task* score:

$$\text{AGENTIC} = \sum_{k=1}^5 w_k s_k, \quad \sum_{k=1}^5 w_k = 1, \quad (5)$$

$$\text{OVERALL} = \frac{1}{2} \text{AGENTIC} + \frac{1}{2} \text{TASK}. \quad (6)$$

Weights emphasize the trust-critical stages, with grounding (s_2) assigned the highest weight and symbolic verification and refinement the second highest; our default is $\mathbf{w} = (w_1, \dots, w_5) = (0.20, 0.30, 0.20, 0.20, 0.10)$.

Experiments

Datasets and Tasks

ARISE is evaluated across two IBD imaging domains: Crohn’s disease (CD) using wireless capsule endoscopy (WCE) and ulcerative colitis (UC) using B-mode intestinal ultrasound. Detailed dataset composition, support/few-shot examples/test splits, and task definitions are provided in Appendix B.

For Crohn’s disease, we use the CrohnIPI dataset (Vallée et al. 2020), a wireless capsule endoscopy (WCE) collection of 3,498 RGB images independently reviewed by multiple clinical experts. We evaluate three tasks of increasing diagnostic granularity: (1) Pathological vs non-pathological image classification, (2) ulceration diagnosis (ulceration vs. no ulceration), and (3) six-lesion classification for pathological images, with labels being aphthoid ulceration, stenosis, edema, erythema, ulceration between 3 and 10 mm, and ulceration > 10 mm.

For Ulcerative Colitis, ultrasound data is taken from the C-TRUS dataset (Leenings et al. 2025), containing a total of 800 ultrasound images collected from 13 patients. We evaluate colon-wall stratification status, classifying whether bowel-wall layering is preserved or lost, a clinically relevant marker of intestinal inflammation.

Models and Prompting

We evaluate ARISE using instruction-following VLMs from three model families: GPT, Gemini, and Qwen. We test GPT-5.4 (Afridi et al. 2026), GPT-4o-mini (Achiam et al. 2023), Gemini-2.5 (Comanici et al. 2025), and Qwen2-VL (Wang et al. 2024), with Qwen2-VL serving as an open-weight VLM baseline. All models are evaluated on identical query splits. Automatic hypothesis generation and logical premise construction use a reasoning-oriented language model, while symbolic contradiction detection is performed using the Z3 SMT solver (De Moura and Bjørner 2008).

We perform 2 major experiments. 1. We compare the diagnostic accuracy, F1 score, precision, and recall of the base VLM and the VLM with the final generated hypotheses (task score after stage 5). 2. We then compare the agentic scores for every stage across models, and the overall task score for every model, on the 6-lesion classification task. This would tell us which stages are strongest and weakest, and which models tend to outperform the others in both workflow and

Table 1: Base and ARISE classification accuracy across the three Crohn’s disease WCE tasks. Task 1: Pathological vs non-pathological image classification, Task 2: Ulceration diagnosis, Task 3: 6-lesion classification

Model	Task 1	Task 2	Task 3
GPT-5.4 (Base)	0.69	0.66	0.32
GPT-5.4 (ARISE)	0.88	0.83	0.79
GPT-4o-mini (Base)	0.61	0.61	0.29
GPT-4o-mini (ARISE)	0.88	0.81	0.74
Gemini-2.5 (Base)	0.66	0.33	0.37
Gemini-2.5 (ARISE)	0.85	0.71	0.58
Qwen2-VL (Base)	0.42	0.66	0.22
Qwen2-VL (ARISE)	0.83	0.74	0.64

Table 2: Colon wall stratification status classification

Model	Accuracy	F1	Precision	Recall
GPT-5.4 (Base)	0.60	0.66	0.69	0.63
GPT-5.4 (ARISE)	0.87	0.78	0.83	0.74
GPT-4o-mini (Base)	0.59	0.58	0.58	0.47
GPT-4o-mini (ARISE)	0.77	0.75	0.76	0.74
Gemini-2.5 (Base)	0.26	0.37	0.63	0.22
Gemini-2.5 (ARISE)	0.71	0.78	0.76	0.72
Qwen2-VL (Base)	0.66	0.64	0.63	0.64
Qwen2-VL (ARISE)	0.74	0.74	0.75	0.76

task-level performance. Crucially, these evaluations jointly establish the clinical value of verifiable reasoning over standard prompting, while opening the black box to reveal exactly how process-level reasoning quality drives final diagnostic accuracy.

Results

Diagnostic Performance: Base VLM vs. ARISE

Table 1 compares base-model accuracy with the same VLMs operating within the ARISE workflow across the three Crohn’s disease WCE tasks. ARISE improves accuracy for every model, with the largest accuracy gain observed as the task complexity increases (6-lesion classification being the hardest) and for the weaker baseline models. Qwen2-VL and Gemini-2.5 close much of the gap to the stronger GPT baselines once wrapped in the ARISE workflow, suggesting that the framework’s gains come from structuring and verifying reasoning rather than from any one backbone’s raw visual capability. Detailed task-level results for this dataset, including accuracy, F1 score, precision, and recall for all model–task combinations, are provided in the Appendix.

For the ulcerative-colitis cohort, Table 2 reports the corresponding accuracy, F1, precision, and recall for the B-mode ultrasound colon-wall-stratification task (preserved vs. lost). The ARISE framework results in a consistent improvement over all baseline models even when diagnosing from noisy B-mode ultrasound images, showing that the performance improvement generalizes across different tasks and modalities.

Stage-Level Reasoning Quality

Figure 3 reports the five individual stage scores (s1 through s5) for all four models on the six-lesion classification task, together with the mean across models at each stage. Hypothesis generation (S1) and symbolic verification (S4) are consistently strong across models, while evidence-guided refinement (S3) is the most challenging stage and exhibits the greatest model variation. The contrast between the two kinds of stages points to a broader pattern: agents are consistently better at rule-based steps than at visually-grounded refinement, where they must revisit the image evidence itself and use it to revise their own hypotheses rather than follow a fixed procedure. The strongest models also perform better during refinement and final diagnosis, indicating that visually grounded adaptation is a key factor separating stronger from weaker agents.

Agentic and Task-level scores across models

We examine how each model’s intermediate reasoning quality (the agentic score) relates to its final classification performance (the task score) on the six-lesion classification task. Figure 4 reports the overall, agentic, and task scores for all four models, ranked by overall score.

Across models, stronger agentic performance generally corresponds to stronger diagnostic performance, but the relationship is not one-to-one. The rule-based stages, such as hypothesis generation and symbolic verification, are consistently strong across all models, whereas the largest differences emerge during evidence-guided refinement and final diagnosis, where the model must use visual evidence to revise and apply its hypotheses. Hence weaker models can maintain relatively high agentic scores by successfully completing structured stages such as hypothesis generation and verification, while still struggling to translate that reasoning into the correct final classification. The resulting process–outcome gap highlights the value of ARISE’s stage-level evaluation, which reveals where reasoning quality stops carrying through to diagnostic performance.

Qualitative Visual Grounding

To illustrate the mechanism behind these quantitative gains, Figure 5 presents a representative B-mode ultrasound case from the colon-wall-stratification task with ground-truth label “lost.” Queried directly, GPT-5.4 hedges between competing visual interpretations and concludes an incorrect diagnosis. Under the ARISE workflow, the same underlying model grounds its answer in concrete, vocabulary-anchored observations (“the bowel wall appears diffusely hypoechoic and homogeneous, with no separable alternating layers or consistent banding along the segment”) and correctly predicts “lost.” This case is typical of the pattern underlying the quantitative improvements described earlier: ARISE’s evidence-summarization (S2) and symbolic-verification (S4) stages guide the model away from ambiguous, hedged visual claims and toward grounded and verifiable ones, which is where the framework’s accuracy gains originate.

Path to Deployment

ARISE is designed as a clinician-in-the-loop decision-support system rather than an autonomous diagnostic tool, emphasizing interpretability, error detection, and compatibility with existing IBD imaging workflows. The current framework has been evaluated across two complementary IBD imaging settings: 3,498 wireless capsule endoscopy (WCE) images from the publicly available CrohnIPI dataset (Vallée et al. 2020) and B-mode intestinal ultrasound data obtained from the C-TRUS dataset (Leenings et al. 2025) and annotated by an expert at the Indiana University (IU) School of Medicine. Clinical experts informed the visual cues, decision boundaries, and interpretation guidelines used by ARISE. Unlike black-box deep-learning based models, ARISE exposes the visual findings, diagnostic hypotheses, symbolic verification results, refinement steps, and final recommendation, allowing clinicians to inspect or override reasoning before acting on the output.

The next deployment stage is retrospective clinical validation using anonymized IU School of Medicine data from approximately 40 adults with Crohn’s disease, collected by a gastrointestinal ultrasound expert certified by the International Bowel Ultrasound Group (IBUS), an international organization supporting standardized training, research, and clinical adoption of intestinal ultrasound for IBD care. Expert annotations of bowel-wall stratification and clinical assessments will serve as reference standards for evaluating ARISE predictions. Stage-level errors will be analyzed across disease severity, patient variation, image quality, and acquisition hardware.

For patients with anonymized, clinically indicated endoscopy data, ARISE will also estimate the Simple Endoscopic Score for Crohn’s Disease (SES-CD) (Daperno et al. 2004), from AI-identified ulcer size, ulcerated surface, and luminal narrowing across the terminal ileum and four colonic segments. This will be compared to the SES-CD score assessments by experts.

Following retrospective validation, ARISE will be incorporated into a software-based clinical interface capable of processing endoscopy and B-mode ultrasound data streams. In an initial shadow-mode study, ARISE outputs will be compared with specialist interpretation without influencing patient care. A subsequent clinician-in-the-loop pilot will allow specialists to accept, question, or override intermediate reasoning and final recommendations, providing direct evidence of where the system’s explanations are clinically useful or insufficient.

Finally, nurses and generalist clinicians will participate in a small case study evaluating whether ARISE’s visual-cue explanations improve non-expert interpretation and diagnostic accuracy relative to unaided assessment (Figure 6). Successful prospective evaluation would support testing ARISE across multiple hospitals and integrating it with existing clinical imaging and reporting systems. This staged progression, from retrospective validation, to silent real-time testing, to clinician-assisted evaluation and multi-site deployment, provides a practical path toward safe, interpretable use of ARISE in IBD assessment.

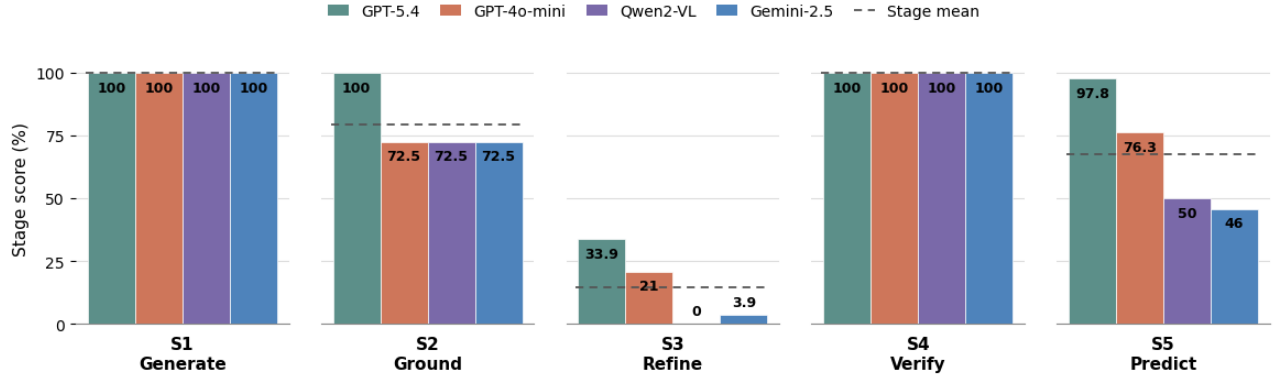


Figure 3: Stage-wise agentic scores across models on the Crohn’s disease 6-lesion classification task for the different workflow stages: hypothesis generation (S1), visual grounding (S2), refinement (S3), symbolic verification (S4), and final diagnosis (S5). Rule-based stages are consistently strong, while evidence-guided refinement shows the greatest difficulty and model variation.

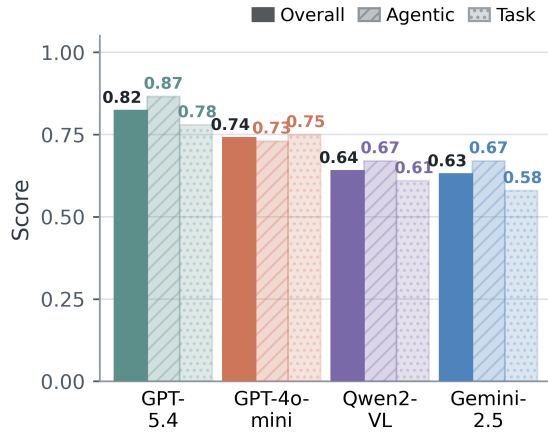


Figure 4: Overall, agentic, and task scores for the 4 evaluated agents on the Crohn’s disease 6-lesion classification task.

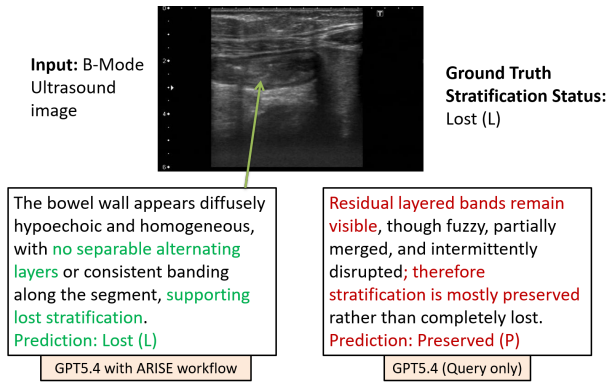


Figure 5: Qualitative example of improvement of visual grounding with ARISE. The base VLM produces ambiguous visual cues, ultimately leading to the wrong classification. ARISE improves visual grounding and ultimately yields the correct classification.

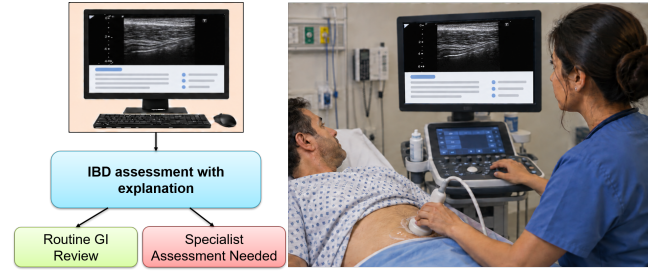


Figure 6: Illustration of an AI-enabled decision-support interface for explainable IBD assessment using intestinal ultrasound, designed to make model outputs accessible and interpretable to nurses.

Conclusion

We presented ARISE, a five-stage agentic framework that adapts a frozen vision-language model to few-shot IBD assessment by generating, grounding, verifying, and refining diagnostic hypotheses without any fine-tuning. Across two independent cohorts (WCE for Crohn’s disease and B-mode ultrasound for ulcerative colitis) and three VLM families, ARISE improved accuracy, F1, precision, and recall over baselines on every model–task combination, with the largest gains on the hardest task and weakest baseline model, while its stage-level scoring exposed cases where strong intermediate reasoning did not fully carry through to final accuracy, highlighting a gap a single outcome metric would hide. Beyond raw accuracy, ARISE’s key contribution is auditability: every reasoning stage is stored as an inspectable artifact rather than collapsed into an opaque prediction, directly addressing the domain-reasoning gap, hallucination risk, data scarcity, and black-box behavior limiting clinical VLM adoption. Future work extends this toward improving the model with the agentic scores and deployment in real-world clinical settings.

References

- Achiam, J.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Afridi, H.; Ullah, H.; Khan, S. D.; and Ullah, M. 2026. From GPT-3 to GPT-5: Mapping their capabilities, scope, limitations, and consequences. *arXiv preprint arXiv:2604.10332*.
- Alayrac, J.-B.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.
- Bots, S.; Nylund, K.; Löwenberg, M.; Gece, K.; Gilja, O. H.; and D’Haens, G. 2018. Ultrasound for assessing disease activity in IBD patients: a systematic review of activity scores. *Journal of Crohn’s and Colitis*, 12(8): 920–929.
- Byun, J. Y.; Park, Y.-J.; Azizan, N.; and Chellappa, R. 2025. Test-Time-Scaling for Zero-Shot Diagnosis with Visual-Language Reasoning. *arXiv preprint arXiv:2506.11166*.
- Byun, J. Y.; Park, Y.-J.; Corbeil, J.-P.; and Abacha, A. B. 2026. Overconfidence and Calibration in Medical VQA: Empirical Findings and Hallucination-Aware Mitigation. *arXiv preprint arXiv:2604.02543*.
- Comanici, G.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv*.
- Daperno, M.; D’Haens, G.; Van Assche, G.; Baert, F.; Bullois, P.; Maunoury, V.; Sostegni, R.; Rocca, R.; Pera, A.; Gevers, A.; Mary, J.-Y.; Colombel, J.-F.; and Rutgeerts, P. 2004. Development and validation of a new, simplified endoscopic activity score for Crohn’s disease: the SES-CD. *Gastrointestinal Endoscopy*, 60(4): 505–512.
- De Moura, L.; and Bjørner, N. 2008. Z3: An efficient SMT solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*, 337–340. Springer.
- Du, Y.; Yu, N.; and Wang, S. 2025. Medical Knowledge Intervention Prompt Tuning for Medical Image Classification. *IEEE Transactions on Medical Imaging*, 44(12): 4945–4959.
- Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; Ji, R.; Shan, C.; and He, R. 2025. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. In *Advances in Neural Information Processing Systems*, volume 38. Datasets and Benchmarks Track.
- Garcez, A. d.; and Lamb, L. C. 2023. Neurosymbolic ai: The 3rd wave. *Artificial Intelligence Review*, 56(11): 12387–12406.
- Hager, P.; et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9): 2613–2622.
- He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations (ICLR)*.
- Iddan, G.; Meron, G.; Glukhovsky, A.; and Swain, P. 2000. Wireless capsule endoscopy. *Nature*, 405(6785): 417.
- Ji, Z.; et al. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.
- Jiang, Y.; et al. 2025. MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents. *NEJM AI*, 2(9): AIdbp2500144.
- Kaplan, G. G.; and Windsor, J. W. 2021. The four epidemiological stages in the global evolution of inflammatory bowel disease. *Nature Reviews Gastroenterology & Hepatology*, 18(1): 56–66.
- Leenings, R.; Konowski, M.; Winter, N. R.; Ernsting, J.; Fisch, L.; Barkhau, C.; Dannlowski, U.; Lügering, A.; Jiang, X.; and Hahn, T. 2025. C-TRUS: A Novel Dataset and Initial Benchmark for Colon Wall Segmentation in Transabdominal Ultrasound. In *Simplifying Medical Ultrasound*, 101–111. Springer Nature Switzerland.
- Li, L.; Qu, J.; Song, L.; Zhou, Y.; Qin, Y.; Yang, T.; and Zhao, Y. 2025. Treble Counterfactual VLMs: A Causal Approach to Hallucination. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 18423–18434. Association for Computational Linguistics.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Liu, Y.; et al. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.
- Maaser, C.; et al. 2020. Intestinal ultrasound for monitoring therapeutic response in patients with ulcerative colitis: results from the TRUST&UC study. *Gut*, 69(9): 1629–1636.
- Madaan, A.; et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36: 46534–46594.
- Ng, S. C.; et al. 2017. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *The Lancet*, 390(10114): 2769–2778.
- Olausson, T.; et al. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5153–5176.
- Pan, L.; Albalak, A.; Wang, X.; and Wang, W. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3806–3824.
- Radford, A.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Ram, O.; et al. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11: 1316–1331.
- Schmidgall, S.; Ziaei, R.; Harris, C.; Reis, E.; Jopling, J.; and Moor, M. 2024. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Shakeri, F.; et al. 2024. Few-shot adaptation of medical vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 553–563. Springer.

Shu, M.; et al. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35: 14274–14289.

Torres, J.; Mehandru, S.; Colombel, J.-F.; and Peyrin-Biroulet, L. 2017. Crohn’s disease. *The Lancet*, 389(10080): 1741–1755.

Turner, D.; et al. 2021. STRIDE-II: An update on the Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE) initiative of the International Organization for the Study of IBD (IOIBD). *Gastroenterology*, 160(5): 1570–1583.

Vallée, R.; et al. 2020. CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn’s disease lesions recognition algorithms. In *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 11317, 440–446. SPIE.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution. arXiv:2409.12191.

Xia, P.; Zhu, K.; Li, H.; Wang, T.; Shi, W.; Wang, S.; Zhang, L.; Zou, J.; and Yao, H. 2025. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. In *International Conference on Learning Representations*.

Zhan, Z.; Zhou, S.; Zhou, X.; Xiao, Y.; Wang, J.; Deng, J.; Zhu, H.; Hou, Y.; Song, Y.; Lin, M.; and Zhang, R. 2026. Retrieval-augmented in-context learning for multi-modal large language models in disease classification. *Journal of Biomedical Informatics*, 178: 105017.

Zhang, J.; Wang, G.; Kalra, M. K.; and Yan, P. 2025. Disease-informed adaptation of vision-language models. *IEEE Transactions on Medical Imaging*, 44(10): 3984–3996.

A. Stage Scoring Definitions and Equations

Notation. Let $\mathcal{Y}$ denote the task label space, $H^{(0)}$ the initial hypothesis set, $\mathcal{E}^*$ the accumulated image-grounded evidence, H^* the final verified hypothesis set, $\mathcal{V}$ the controlled clinical vocabulary, Γ the natural-language-to-first-order-logic parser, and $\Phi = \Gamma(H)$ the corresponding logical premises. $\hat{y}$ denotes the predicted label and $\hat{h}$ the hypothesis cited in support of that prediction.

Score s_1 . $S1$ is scored by a three-item rubric averaged to $[0, 1]$:

$$s_1 = \frac{1}{2} \left(\underbrace{\mathbf{1}[H^{(0)} \neq \emptyset]}_{\text{non-empty}} + \underbrace{\frac{\sum_{y \in \mathcal{Y}} \mathbf{1}[\exists h \in H^{(0)}, x \rightarrow^h y]}{|\mathcal{Y}|}}_{\text{per-label coverage}} \right) \quad (\text{A1})$$

where the first term indicates that the hypotheses set generated has at least 1 hypothesis, and the 2nd term indicates that there is at least one hypothesis per label. The best agentic score ($s_1 = 1$) is obtained when both conditions hold.

Score s_2 . Let $\mathcal{F}(e_t)$ be the findings cited by the agent and $\mathcal{F}^*(x_t)$ the label-relevant findings marked present by expert annotation of the (few) support images. We report precision and recall separately and combine them, weighting fidelity above completeness since a fabricated finding is the clinical dealbreaker:

$$\text{Prec} = \frac{|\mathcal{F}(e_t) \cap \mathcal{F}^*(x_t)|}{|\mathcal{F}(e_t)|}, \quad \text{Rec} = \frac{|\mathcal{F}(e_t) \cap \mathcal{F}^*(x_t)|}{|\mathcal{F}^*(x_t)|}, \quad (\text{A2})$$

$$\text{Obs} = \frac{|\{c \in e_t : c \in \mathcal{V}\}|}{|e_t|}, \quad (\text{A3})$$

$$s_2 = 0.5\text{Prec} + 0.4\text{Rec} + 0.1\text{Obs} \quad (\text{A4})$$

where Obs (observability) is the fraction of claims referencing a concrete vocabulary finding rather than an abstract inference.

Score s_3 . Refinement is scored by the improvement in prediction accuracy of downstream task before and after the refinement step, when compared against the ground truth label, for a smaller validation set of 50 observations.

$$s_3 = \text{Acc}_{H(\mathcal{E}^*)} - \text{Acc}_{H^{(0)}} \quad (\text{A5})$$

Score s_4 . We score the translation fidelity of the LLM parser using a Natural Language Inference (NLI) model (DeBERTa-v3-large) (He et al. 2021). First, we compute the natural language hypothesis back from the first order logic:

$$\tilde{H}^* = \Gamma^{-1}(\Phi) \quad (\text{A6})$$

We then test whether H^* and $\tilde{H}^*$ mutually entail one another (each statement implies the other). The NLI gives us the entailment probability (probability of one statement implying the other) and we get the score as the min of 2 probabilities.

$$s_4 = \min \left[P_e(H^* \rightarrow \tilde{H}^*), P_e(\tilde{H}^* \rightarrow H^*) \right] \quad (\text{A7})$$

This ensures that the 2 statements are equivalent and the first order logic statement was correct.

S5: Final diagnosis. The process score checks whether the prediction lies in the permitted label space and whether the model cites a hypothesis from the verified set as support:

$$s_5 = \frac{1}{2} \left[\mathbf{1}(\hat{y} \in \mathcal{Y}) + \mathbf{1}(\hat{h} \in H^*) \right]. \quad (\text{A8})$$

Task score. Diagnostic outcome quality is measured separately from process quality as the mean of accuracy, F1 score, precision, and recall on the held-out query set:

$$\text{TASK} = \frac{1}{4} [\text{Acc} + \text{F1} + \text{Prec} + \text{Rec}]. \quad (\text{A9})$$

Workflow-level aggregation. The five stage scores are combined using the weights reported in the main manuscript, with the greatest weight assigned to image grounding:

$$\text{AGENTIC} = \sum_{k=1}^5 w_k s_k, \quad (\text{A10})$$

$$\text{OVERALL} = \frac{1}{2} \text{AGENTIC} + \frac{1}{2} \text{TASK}. \quad (\text{A11})$$

where $(w_1, \dots, w_5) = (0.20, 0.30, 0.20, 0.20, 0.10)$

B. Datasets and experimental protocol

Crohn’s disease (Wireless Capsule Endoscopy) The CrohnIPI dataset (Vallée et al. 2020) contains 3,498 expert-reviewed WCE RGB images. The manuscript reports 60.55% normal and 38.85% pathological images. Pathological images include erythema, edema, aphthoid ulceration, ulceration between 3 and 10mm, ulceration greater than 10mm, and stenosis; some images contain co-occurring labels.

Ulcerative colitis (B-Mode Ultrasound) The C-TRUS dataset (Leenings et al. 2025) contains 800 B-mode ultrasound images from 13 patients. The study filters 275 good-quality images for the colon-wall stratification experiment. These raw datapoints were annotated for colon-wall stratification classification, and segmentation of the colon wall, by an expert at IU School of Medicine.

The datasets, tasks and number of support and in-context examples and number of test samples are mentioned in Table A1

Experimental Protocol

1. **Base VLM evaluation:** Each backbone is first evaluated in a query-only setting on the held-out test split.
2. **ARISE evaluation:** The same backbone is evaluated on the same query split after the five-stage ARISE workflow constructs the verified hypothesis state H^* .
3. **Task-level comparison:** Accuracy, F1 score, precision, and recall are reported for Base and ARISE on all four clinical tasks.
4. **Stage-level analysis:** Scores S1–S5 are compared across the four VLMs on the Crohn’s disease six-lesion classification task. S3 is computed on a separate 50-example validation subset.
5. **Process vs. outcome analysis:** The weighted AGENTIC score is compared with the TASK score to identify cases in which a well-formed reasoning process does not fully translate to correct final classification.

C. Results for Crohn’s disease

Tables A2, A3 and A4 below report the full task-level metrics for each backbone in the query-only Base setting and within ARISE.

D. Prompt Templates and Implementation

Prompt A: Normal vs. Pathological WCE Classification

The binary prompt assigns each WCE frame to exactly one of two labels, N (normal/non-pathological) or P (pathological). The prompt defines pathological findings to include aphthoid ulceration, ulceration, erythema, edema, stenosis, focal mucosal breaks, depressed lesions, white/yellow ulcer bases, red halos, abnormal redness, swelling, or luminal narrowing. The decision rules instruct the model to choose N only when the mucosa appears continuous and no focal abnormality is visible, and to choose P whenever a definite abnormality is present. In the baseline setting, no refined hypotheses or in-context examples are provided. In the ARISE setting, the refined hypothesis list is inserted under a *Supporting rules* field and the model is asked to return the predicted class together with the ID of the supporting hypothesis when available.

Prompt A template:

You are a medical assistant specialized in wireless capsule endoscopy image classification.

Classify each WCE image into exactly one binary label:

N = normal-appearing / non-pathological frame.

P = pathological / abnormal frame.

Supporting rules: [refined hypotheses, ARISE setting only]

Decision rules: choose N only for normal, continuous mucosa without focal abnormality; choose P if any visible abnormality is present; do not output a lesion subtype.

Prompt B: Six-Lesion WCE Classification

Prompt B is applied only to frames already considered pathological and assigns exactly one lesion label: aphthoid ulceration (AU), erythema (E), edema (O), stenosis (S), ulceration 3–10 mm (U3–10), or ulceration >10 mm (U>10). The prompt includes a short description of each class and an ulcer-first decision hierarchy: visible mucosal break, depression, erosion, crater, or white/yellow ulcer base directs the model toward an ulcer subtype; intact-surface redness favors erythema; swollen/enlarged villi with a continuous surface favors edema; and dominant luminal narrowing favors stenosis. The baseline Prompt B uses a fixed set of short class-discrimination rules. In the ARISE setting, these fixed rules are replaced by the refined hypothesis state loaded from disk.

Prompt B template:

You are a medical assistant specialized in wireless capsule endoscopy image classification.

This image is already considered pathological. Do not choose N.

Classify the pathological subtype into exactly one of:

AU, E, O, S, U3–10, U>10.

Supporting rules: [fixed short rules for baseline; refined hypotheses for ARISE]

Decision hierarchy: ulcerative surface loss → ulcer subtype; intact-surface redness → E; swollen/enlarged villi with continuous surface → O; dominant luminal narrowing → S.

Table A1: Dataset and task splits used in the experiments. AU: aphthoid ulceration; S: stenosis; O: edema; E: erythema; U3-10: ulceration between 3 and 10 mm; U>10: ulceration > 10 mm.

Task	Modality	Output labels	Support	ICL	Test
Normal vs. pathological classification	WCE	Normal; pathological	7	18	1434
Six-lesion classification	WCE	AU; S; O; E; U3-10; U>10	7	12	535
Ulceration diagnosis	WCE	Ulceration; no ulceration	7	12	535
Colon-wall stratification status	B-mode ultrasound	Preserved; lost	4	10	211

Table A2: Pathological vs Non-Pathological WCE image

Model	Accuracy	F1	Precision	Recall
GPT-5.4 (Base)	0.69	0.65	0.59	0.72
GPT-5.4 (ARISE)	0.88	0.86	0.85	0.89
GPT-4o-mini (Base)	0.61	0.73	0.60	0.77
GPT-4o-mini (ARISE)	0.88	0.88	0.89	0.88
Gemini-2.5 (Base)	0.66	0.59	0.77	0.62
Gemini-2.5 (ARISE)	0.85	0.78	0.82	0.76
Qwen2-VL (Base)	0.42	0.29	0.21	0.50
Qwen2-VL (ARISE)	0.83	0.83	0.88	0.83

up to 12 examples, Gemini uses four, and Qwen2-VL-2B uses two. Qwen uses fewer examples to reduce GPU memory consumption on a Colab T4.

Table A3: WCE - Diagnosing ulceration

Model	Accuracy	F1	Precision	Recall
GPT-5.4 (Base)	0.66	0.76	0.79	0.73
GPT-5.4 (ARISE)	0.83	0.78	0.83	0.74
GPT-4o-mini (Base)	0.61	0.67	0.86	0.59
GPT-4o-mini (ARISE)	0.81	0.84	0.88	0.81
Gemini-2.5 (Base)	0.33	0.18	0.78	0.10
Gemini-2.5 (ARISE)	0.71	0.78	0.86	0.72
Qwen2-VL (Base)	0.66	0.77	0.76	0.79
Qwen2-VL (ARISE)	0.74	0.84	0.75	0.96

Support and In-Context Examples

A set of support examples are selected for each task for Stage S2. Separately, ARISE runs use manually selected labeled image examples as multimodal in-context demonstrations. The demonstrations contain the example image and its class label; the explanatory text stored with each example is used during the final diagnostic evaluation. For Prompt A, GPT and Gemini use four ICL examples, while Qwen2-VL-2B uses two. For Prompt B, the GPT notebook permits

Table A4: 6 lesion classification from Crohn’s disease WCE

Model	Accuracy	F1	Precision	Recall
GPT-5.4 (Base)	0.32	0.30	0.35	0.32
GPT-5.4 (ARISE)	0.79	0.76	0.81	0.75
GPT-4o-mini (Base)	0.29	0.34	0.32	0.26
GPT-4o-mini(ARISE)	0.74	0.75	0.77	0.74
Gemini-2.5 (Base)	0.37	0.35	0.47	0.37
Gemini-2.5(ARISE)	0.58	0.53	0.58	0.63
Qwen2-VL (Base)	0.22	0.11	0.07	0.22
Qwen2-VL (ARISE)	0.64	0.58	0.58	0.64