Boundary-Guided Multimodal Reasoning for Explainable Gastrointestinal VQA

Tien-Dat Dam[1,2], Le-Tran Nguyen[1,2] and Trung-Nghia Le[1,2,†]

[1]University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

The clinical adoption of artificial intelligence in gastroenterology remains limited by the lack of interpretability in existing Visual Question Answering (VQA) systems. Although recent self-probing approaches can generate textual explanations, they often lack explicit grounding in relevant visual evidence. To address this limitation, we propose a boundary-guided multimodal reasoning framework for explainable gastrointestinal VQA. We employ a YOLO26-seg model to localize precise object boundaries around polyps and clinical instruments, preserving internal textures while highlighting diagnostically relevant regions. A fine-tuned Qwen3.5-4B model leverages these boundary-enhanced images to perform progressive visual reasoning, while a two-stage self-probing strategy combined with a Qwen3.5-35B-A3B explanation module produces coherent clinical justifications for the final predictions. Experiments on the Kvasir-VQA-x1 dataset demonstrate that the proposed framework improves both VQA performance and explanation quality compared with baseline methods and conventional visual grounding strategies, offering a more trustworthy approach to gastrointestinal decision support.

1. Introduction

Medical Visual Question Answering (VQA) has emerged as a promising approach for developing clinical decision support systems in endoscopy [1, 2, 3]. By leveraging vision-language models (VLMs), medical VQA combines natural language processing with image understanding to answer complex diagnostic questions [4, 5, 6]. Recent benchmark initiatives, such as ImageCLEF 2026 [7] and the MediaEval Medico 2026 Challenge [8], underscore a growing demand for multimodal AI systems that not only deliver high diagnostic accuracy but also provide interpretable clinical support. Specialized datasets, such as Kvasir-VQA [9], further support this goal by providing dedicated benchmarks for reasoning within gastrointestinal imaging.

Despite the potential of VQA models, their integration into real-world clinical workflows is severely hindered by their "black-box" nature, which lacks the interpretability required for trustworthy medical diagnosis [10, 11]. While recent methods [12] have utilized prompt-based self-probing to generate textual justifications, they suffer from a significant limitation: a lack of explicit visual grounding. Without the direct localization of anatomical landmarks or pathologies, text-only reasoning often creates a disconnect where generated explanations fail to accurately reflect the underlying visual evidence, ultimately reducing clinical trust in the AI's predictions.

To overcome these limitations, we propose a boundary-guided multimodal reasoning framework designed to enhance explainable gastrointestinal VQA. Our framework uses a YOLO26-seg

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
*Corresponding author.
23120118@student.hcmus.edu.vn (T. Dam); nltran2525@apcs.fitus.edu.vn (L. Nguyen); ltnghia@fit.hcmus.edu.vn (T. Le)
0009-0009-9507-2392 (T. Dam); 0009-0009-0736-6244 (L. Nguyen); 0000-0002-7363-2610 (T. Le)

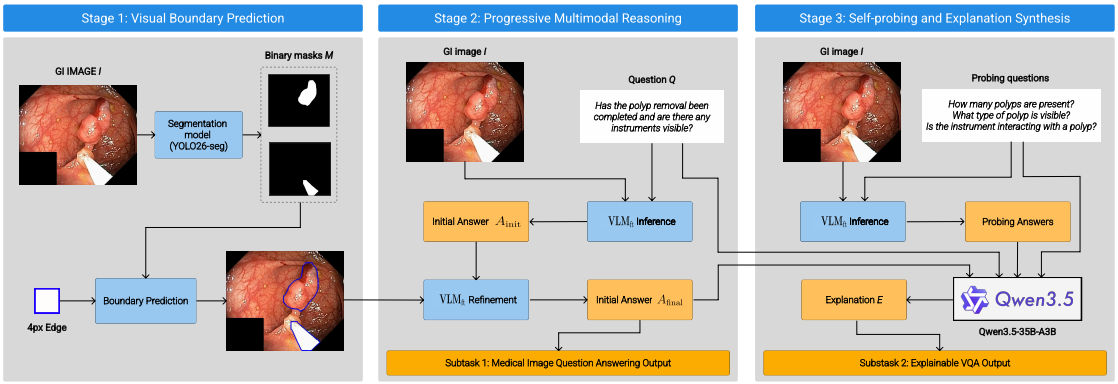

Figure 1: Overview of the explainable gastrointestinal VQA pipeline, featuring: (1) visual boundary prediction for texture-preserving localization; (2) interactive multimodal reasoning to refine diagnostic predictions; and (3) explanation synthesis to output transparent, visually-grounded clinical reasoning.

model [13], which generates precise 4-pixel blue boundaries around regions of interest (*e.g.*, polyps or clinical instruments). This approach preserves the internal textures of GI structures while offering sharp spatial localization. We also design a progressive inference pipeline where a fine-tuned Qwen3.5-4B model [14] generates an initial prediction and subsequently refines it using the boundary-guided image. Finally, we leverage a comprehensive self-probing strategy coupled with a powerful Qwen3.5-35B-A3B [14] explanation synthesizer to generate well-grounded, highly structured clinical explanations. Experimental evaluations on the Kvasir-VQA-x1 dataset [15] demonstrate that our proposed framework improves both overall VQA performance and the quality of clinical explanations compared to state-of-the-art baseline models and conventional visual grounding strategies.

Our contributions are as follows:

- We introduce a novel explainable medical VQA framework that resolves the lack of visual grounding in existing text-based models by introducing a visual boundary guidance module. YOLO26-seg model is utilized to generate sharp, 4-pixel boundaries to accurately localize regions of interest, such as polyps and instruments, without obscuring the internal textures necessary for clinical evaluation.
- We propose a progressive inference pipeline in which a fine-tuned VLM systematically refines its predictions using boundary-guided images to enhance diagnostic accuracy and visual grounding.
- We develop a self-probing mechanism with a Qwen3.5-35B-A3B synthesizer to generate visually grounded, structured clinical explanations from targeted image queries.

2. Proposed Method

2.1. Overview

Our framework employs a dual-model architecture that combines high VQA performance with clinician-oriented explainability through a three-stage pipeline (Fig. 1). First, Visual Boundary Prediction uses a segmentation model to delineate polyps and clinical instruments, producing boundary-guided images that preserve diagnostically relevant details. Second, during Progressive Multimodal Reasoning, a fine-tuned Qwen3.5 model iteratively refines its predictions using these boundary cues. Finally, Self-Probing and Explanation Synthesis leverages a Qwen3.5 model to integrate the refined prediction with responses to targeted probing questions, generating coherent clinical explanations.

2.2. Visual Boundary Prediction

Traditional visual grounding methods like bounding boxes and segmentation masks have severe limitations in clinical settings. Bounding boxes capture redundant background information, while segmentation masks completely obscure the underlying pathology or instrument. To resolve this, we introduce a visual boundary approach. Specifically, we train YOLO26 segmentation models [13] on the Kvasir-SEG [16] and Kvasir-Instrument [17] datasets. By rendering the outer mask contour as a thin, 4-pixel blue line, we precisely localize target structures without obscuring critical internal clinical textures.

Given an input image I, the model predicts a binary segmentation mask $M \in \{0,1\}^{H \times W}$. The boundary image I_b is formulated as: $I_b = I \odot \overline{\partial M} + C_{\text{blue}} \odot \partial M$, where ∂M represents the 4-pixel edge contour of M, $\overline{\partial M}$ present its complement, and C_{blue} is the RGB representation of the blue color. This boundary is dynamically generated for six question classes: *polyp_count, polyp_size, polyp_type, instrument_count, instrument_location, and instrument_presence.*

2.3. Boundary-Guided Reasoning and Self-Probing Explanation

To answer the clinical question Q on image I, our fine-tuned model VLM_{ft} first generates an initial prediction: $A_{\text{init}} = \text{VLM}_{\text{ft}}(I, Q)$. For boundary-relevant questions, the VLM is then prompted with the boundary-overlaid image I_b and the initial prediction to explicitly refine its judgment: $A_{\text{final}} = \text{VLM}_{\text{ft}}(I_b, Q, A_{\text{init}}, \text{Prompt}_{\text{refine}})$. In Subtask 1, A_{final} is directly submitted as the predicted answer.

In Subtask 2, we implement a self-probing mechanism. We map the incoming Q into one of 18 distinct *'question_class'* categories. Each category triggers 10 pre-defined questions $\{q_i\}_{i=1}^{10}$ (*e.g.*, querying shape, size, color, and location) designed to force the model to analyze the image details. The VLM answers each follow-up question: $a_i = \text{VLM}_{\text{ft}}(I, q_i), \quad \forall i \in \{1, \ldots, 10\}$. Finally, the original question Q, the primary answer A_{final}, and the 10 probing pairs $\{(q_i, a_i)\}_{i=1}^{10}$ are fed to the Qwen3.5-35B-A3B model VLM_{syn} to synthesize these elements into a single coherent, structured clinical explanation: $E = \text{VLM}_{\text{sync}}\left(I_b, Q, A_{\text{final}}, \{(q_i, a_i)\}_{i=1}^{10}, \text{Prompt}_{\text{syn}}\right)$.

3. Experiments

3.1. Experimental Settings

Our experiments were conducted on the Kvasir-VQA-x1 dataset [15], which consists of $159,484$ VQA pairs built upon $6,500$ GI endoscopy images from the HyperKvasir and Kvasir-Instrument repositories. We validated methods using the full test split of $15,955$ samples.

VLM models were trained for 1 epoch on the 95% training split of Kvasir-VQA-x1 dataset [15], reserving the remaining 5% for validation. Training was conducted on two NVIDIA A100 GPUs with 40GB VRAM, using a batch size of 128. We applied LoRA to all attention projection layers, MLP blocks, vision and language projections. The LoRA parameters were configured with a rank $r = 16$ and $\alpha = 32$.

3.2. Results

Results in Table 1 highlight the effectiveness of the proposed method. The Qwen3.5-4B backbone consistently outperformed other state-of-the-art models, achieving top scores of 0.4964 in BLEU, 0.7005 in ROUGE-L, and 0.7040 in METEOR. Furthermore, ablation studies comparing different visual grounding techniques demonstrate that the boundary guidance strategy yields a slight

Table 1
Quantitative evaluation of our method against state-of-the-arts on the Kvasir-VQA-x1 dataset. Best and second best results are shown in **bold** and underline.

Model Name	BLEU ↑	ROUGE-L ↑	METEOR ↑
State-of-the-Arts			
Florence-2-base	0.4480	0.6710	0.6744
PaliGemma-3B-pt-224	0.4707	0.6723	0.6744
Qwen3-VL-4B-Instruct	0.4837	0.6914	0.6948
Gemma 4 E4B	0.1581	0.4012	0.4316
Baselines			
Qwen3.5-4B	**0.4964**	0.7005	**0.7040**
Qwen3.5-4B + Bounding Boxes	0.4870	0.6922	0.6972
Qwen3.5-4B + Segmentation Masks	0.4879	0.6924	0.6970
Qwen3.5-4B + Boundary Guidance (Ours)	0.4946	**0.7021**	**0.7040**

but notable improvement over both standard baseline models and traditional visual grounding methods like bounding boxes and segmentation masks. This confirms our hypothesis that preserving the internal textures of clinical images, rather than obscuring them, is vital for accurate clinical reasoning.

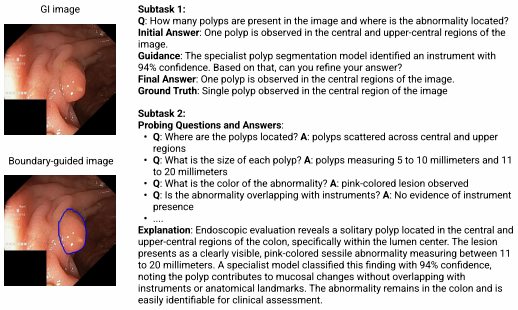

Figure 2: Qualitative example of our method, illustrating answer refinement through model guidance (Subtask 1) and explanation generation via self-probing (Subtask 2).

Fig. 2 provides a qualitative example of the framework's two main subtasks using a gastrointestinal polyp image. Specifically, it illustrates how an initial diagnostic prediction is refined using a 4-pixel visual boundary in Subtask 1, and demonstrates how targeted probing questions are synthesized into a structured clinical explanation in Subtask 2.

4. Conclusion

In this paper, we presented an explainable medical VQA framework for the MediaEval Medico 2026 Challenge. Our dual-system pipeline uses a YOLO26-seg boundary generator, a fine-tuned Qwen3.5-4B VLM, and a Qwen3.5-35B-A3B explanation synthesizer to generate a comprehensive explanation. Experiments show that visual boundaries are clinically superior to coarse bounding boxes and segmentation masks. In future work, we plan to implement an end-to-end training architecture to simultaneously optimize the boundary extraction and the VLM textual reasoning, further reducing inference latency and optimizing clinical integration.

Declaration on Generative AI. We affirm that all ideas, analyses, figures, and conclusions were produced entirely by the human authors. A large language model (LLM) was used only to improve the clarity and readability of the manuscript. No AI tool contributed to the generation of ideas, analysis, or interpretation of results.

References

[1] J. Lau, S. Gayen, A. Ben Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, Scientific Data 5 (2018) 180251. doi:10.1038/sdata.2018.251.

[2] X. He, Y. Zhang, L. Mou, E. Xing, P. Xie, Pathvqa: 30000+ questions for medical visual question answering, arXiv preprint arXiv:2003.10286 (2020).

[3] S. Gaihre, A. T. Magar, P. Pokharel, L. Tiwari, Multimodal ai for gastrointestinal diagnostics: Tackling vqa in medvqa-gi 2025, arXiv preprint arXiv:2507.14544 (2025).

[4] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, 2022, pp. 12888–12900.

[5] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, L. Yuan, Florence-2: Advancing a unified representation for a variety of vision tasks, in: Conference on Computer Vision and Pattern Recognition, 2024, pp. 4818–4829.

[6] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, et al., Paligemma: A versatile 3b vlm for transfer, arXiv preprint arXiv:2407.07726 (2024).

[7] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, International Conference of the CLEF Association, 2026.

[8] S. Gautam, V. Thambawita, M. Riegler, et al., Medico 2026: Visual Question Answering for Gastrointestinal Imaging, arXiv (2026). To be published.

[9] S. Gautam, A. M. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, M. A. Riegler, Kvasir-vqa: A text-image pair gi tract dataset, in: International Workshop on Vision-Language Models for Biomedical Applications, 2024, pp. 3–12.

[10] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, P. Lambin, Transparency of deep neural networks for medical image analysis: A review of interpretability methods, Computers in biology and medicine 140 (2022) 105111.

[11] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, F. Nensa, Explainable ai in medical imaging: An overview for clinical practitioners–beyond saliency-based xai approaches, European journal of radiology 162 (2023) 110786.

[12] S. Gaihre, A. T. Magar, From answers to explanations: Self-probing efficiently fine-tuned vision–language models for medical vqa at medico 2025 (2025).

[13] G. Jocher, J. Qiu, M. Liu, S. Lyu, F. C. Akyon, M. E. Kalfaoglu, Ultralytics yolo26: Unified real-time end-to-end vision models, arXiv preprint arXiv:2606.03748 (2026).

[14] Qwen Team, Qwen3.5: Towards native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.

[15] S. Gautam, M. Riegler, P. Halvorsen, Kvasir-vqa-x1: A multimodal dataset for medical reasoning and robust medvqa in gastrointestinal endoscopy, in: MICCAI Workshop on Data Engineering in Medical Imaging, 2025, pp. 53–63.

[16] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: International conference on multimedia modeling, 2019, pp. 451–462.

[17] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. De Lange, P. T. Schmidt, H. D. Johansen, et al., Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, in: International Conference on Multimedia Modeling, 2021, pp. 218–229.