X-VQA for GI Diagnostics: Multimodal Visual Question Answering with Confidence-Aware Explanations

Krishna Tewari[1,*], Sukomal Pal[1]

[1]Indian Institute of Technology(BHU) Varanasi, India

Abstract

This paper presents a multimodal framework for gastrointestinal (GI) diagnostics, integrating Visual Question Answering (VQA) with structured explanation generation. Leveraging Paligemma-3B and Florence-2 models, the system processes endoscopic images alongside clinician-provided questions to generate accurate answers with clinically grounded explanations. Targeted preprocessing and self-probing mechanisms improve visual-text alignment, while confidence estimation provides insights into model certainty. Evaluated on the Kvasir-VQA-x1 dataset in the Medico 2025 Challenge, the framework demonstrates strong performance across diverse question types, with particularly reliable outputs for queries related to landmarks, instruments, and polyps. The generated explanations are clear, interpretable, and aligned with clinical reasoning, supporting transparent decision-making. Overall, this approach enhances both the accuracy and explainability of AI-assisted GI diagnostics, highlighting the potential of multimodal models to provide clinically meaningful insights in endoscopic image analysis.

1. Introduction and Related Work

Gastrointestinal (GI) diseases constitute a major global health burden, with conditions such as gastric cancer, inflammatory bowel disease, and peptic ulcers affecting millions of individuals. Gastric cancer alone accounts for over one million new cases and approximately 770,000 deaths each year, ranking among the leading causes of cancer mortality worldwide[1]. Timely detection and accurate characterization of GI abnormalities play a critical role in improving patient survival. However, high-resolution endoscopic procedures such as colonoscopy and capsule endoscopy generate extensive volumes of visual data that must be interpreted by experts, making manual assessment labor-intensive, subject to inter-observer variability, and constrained by specialist availability.

Visual Question Answering (VQA) has recently emerged as a promising paradigm for AI-assisted diagnostics in GI imaging, enabling interactive reasoning over complex endoscopic scenes. In this setting, a VQA model takes an endoscopic image I and a clinician-provided question Q to produce an answer A following the mapping $f : (I, Q) \rightarrow A$. Modern transformer-based multimodal architectures jointly encode both visual and textual information to capture subtle relationships between lesions and clinical queries, thereby facilitating detection of early-stage disease and supporting clinical decision-making.

The MediaEval Medico 2025 Challenge [1] specifically emphasizes explainability in GI VQA using the Kvasir-VQA-x1 dataset [2]. Explanation generation improves transparency by producing human-interpretable rationales and confidence estimates, providing clinicians with traceable evidence for model predictions. Earlier medical VQA efforts such as VQA-RAD [3] and PathVQA [4] established foundational benchmarks by pairing medical images with clinically meaningful question–answer pairs and enabling rationale generation. In the GI domain, Kvasir-VQA [5], Kvasir-VQA-x1 [2], and HyperKvasir [6] provide large-scale endoscopic repositories capturing diverse anatomical and pathological variations. Benchmarking efforts such as the ImageCLEFmed-MEDVQA challenges in 2023 [7], 2024 [8, 9] and 2025 [10, 11] have further advanced the field by encouraging innovation in multimodal reasoning on clinical data.

Parameter-efficient tuning methods such as LoRA [12, 13] enable efficient adaptation of large vision–language models, while recent work has demonstrated the utility of Florence2 and CLIPSeg-based architectures in GI VQA with interpretable outputs [14, 15]. Complementary advances in text-

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

*Corresponding author.

krishnatewari.rs.cse24@iitbhu.ac.in (K. Tewari); spal.cse@iitbhu.ac.in (S. Pal)

0009-0005-6599-9956 (K. Tewari); 0000-0001-8743-9830 (S. Pal)

[1]https://www.who.int/news-room/fact-sheets/detail/cancer

conditioned medical image synthesis [16] further support data augmentation and interpretability. Collectively, these developments underscore a clear trend toward explainable, trustworthy, and scalable AI frameworks for GI diagnostics.

2. Methodology

This section presents the methodology for both subtasks of the challenge. Experiments for both of them were conducted on NVIDIA H100 single node with 80GB RAM.

2.1. VQA Pipeline (Subtask 1)

The task demands robust multimodal reasoning that can interpret endoscopic images alongside clinically relevant natural language questions. The dataset presents several unique challenges, including (i) strong illumination artifacts, (ii) circular vignetting effects, and (iii) question diversity requiring compositional understanding. To address these, we designed a two-track methodology utilizing **Paligemma-3B** and **Florence-2-base-ft**, as illustrated in Figure 1.

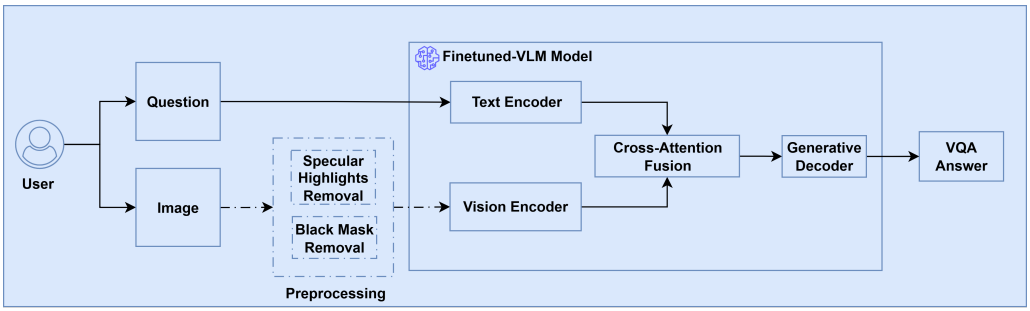


Figure 1: Overview of the proposed methodology for both approaches showing preprocessing.

Preprocessing for Florence-2: Endoscopic images are often affected by optical and environmental distortions that can mislead attention mechanisms in vision-language transformers. Two common artifacts: specular highlights and black masks. They can obscure critical anatomical details, potentially reducing model performance. To address this, a targeted image enhancement module was incorporated into the Florence-2 pipeline. Specular highlights and dark regions are first identified and then corrected using advanced inpainting techniques, which restore continuity in the affected areas while preserving relevant structural information. These preprocessing steps reduce visual noise, normalize illumination, and enhance the alignment of Florence-2's attention mechanisms with diagnostically important features. By improving the clarity and consistency of input images, this module enables more reliable feature extraction and reasoning. In contrast, the Paligemma pipeline demonstrated sufficient robustness to handle raw endoscopic images, so no image enhancement preprocessing was necessary for its operation, allowing it to perform effectively without additional interventions.

Model Training: For the **Paligemma-3B** model[2], parameter-efficient LoRA fine-tuning was applied with rank $r = 8$ and scaling factor $\alpha = 32$, combined with 4-bit NF4 quantization. The visual encoder was frozen, updating only cross-modal projection layers. Training used a learning rate of 1×10^{-5}, mixed precision (`bfloat16`), and gradient accumulation over 16 steps. The **Florence-2-base-ft** model[3] was fine-tuned for three epochs at 3×10^{-5} with batch size three (`fp16`), lora rank $r = 4$, scaling factor $\alpha = 16$ and gradient accumulation over 8 steps. Paligemma outperformed Florence-2 on accuracy and robustness.

2.2. Explanation Generation Pipeline (Subtask 2)

Subtask 2 focuses on generating clinically grounded textual explanations that justify the model's predicted answers. We leveraged the best-performing Subtask 1 backbone, **Paligemma-3B**, enhanced with domain-specific LoRA adapters to maintain consistency with learned visual and linguistic representations.

[2]https://huggingface.co/krissTewari/Kvasir-VQA-x1-lora_pali-gemma
[3]https://huggingface.co/krissTewari/Florence-2-vqa-final

Textual Explanation Generation: Explanations were generated by the fine-tuned Paligemma-3B in instruction-following mode. For each test instance, the pipeline reformulated the question and predicted answer into a structured explanatory prompt of the form: *"Given this endoscopic image and the model's predicted answer, provide a concise, clinically grounded explanation supporting that answer."*

To promote interpretability and structured reasoning, we employed an internal **self-probing mechanism** that decomposes the model's inference into eight targeted diagnostic sub-questions. These were designed to explicitly elicit relevant visual, contextual, and clinical cues before forming the final explanation (What is the abnormality?, Where is it located?, Describe its morphology., Describe its color, size, and vascular pattern., What is the depth or severity of the abnormality?, What is the likely clinical significance or differential diagnosis?, Are there any associated findings?, Suggest possible next steps.)

The model internally answered these, concatenated the latent responses, and synthesized a coherent two-to three-sentence explanation, ensuring factuality and interpretability to VQA outputs.

Confidence Estimation: A quantitative **confidence score** was computed for every generated explanation to represent the model's epistemic certainty. This score was derived from token-level probability distributions within the Paligemma decoder. For each generated token t_i, the model outputs a normalized softmax probability p_i over the vocabulary. The explanation-level confidence C is obtained as the arithmetic mean across all generated tokens:

$$C = \frac{1}{N} \sum_{i=1}^{N} p_i, \tag{1}$$

where N is the total number of generated tokens. The resulting scalar $C \in [0, 1]$ captures the model's internal calibration-values near 1.0 indicate high certainty and linguistic stability, while lower values flag potential ambiguity or weak contextual grounding. The confidence score was stored as part of each explanation output, enabling downstream evaluation of interpretability.

Limitations: This mean-probability surrogate is known to be uncalibrated in absolute terms, as autoregressive likelihoods do not directly reflect epistemic correctness. Furthermore, since C scales with *token-wise* probabilities, it is length-biased: shorter rationales with fewer low-probability tokens tend to inflate C, whereas longer explanations are more likely to include locally uncertain tokens and thus depress the aggregate score, even when semantically correct.

3. Results

This section presents the results for both Subtask 1 and Subtask 2 in the Medico 2025 Challenge. For evaluation official evaluation script provided by organizers' was used. The models were evaluated across the official private and public test sets for Subtask 1 and results are as show in Table 1. The metrics used are ROUGE-1/2/L [17], METEOR [18], BLEU [19], CHRF++ [20] and F1 score wherein higher is better. Figure 2a, Figure 2b represent absolute performance scores of three complexity levels across various question categories for public and private set respectively. The evaluation for Subtask 2 uses the Qwen3-30B-A3B large language model as an automated agent to perform rubric-based semantic assessment of model outputs. Each instance was evaluated across five dimensions: Answer Correctness, Faithfulness, Clinical Relevance, Clarity, and Completeness. The score distribution by question types is shown in Figure 2c. The quantitative results for the same is shown in Table 2.

While our team performed competitively across all metrics for both subtasks, the lower ROUGE-1 score for Subtask 1 arises due to (i) the absence of explicit rationale supervision or ROUGE-focused optimization, (ii) evaluation on raw, unnormalized outputs rather than normalized text, and (iii) the use of constrained PEFT and quantized VLMs. Consequently, this drop reflects stricter evaluation and supervision asymmetry rather than a failure in capturing semantic content.

4. Conclusion and Future Work

In this work, we demonstrated that combining multimodal VQA with structured explanation generation, using models like Paligemma-3B and Florence-2, provides accurate, interpretable, and clinically relevant GI diagnostics. The preprocessing and self-probing mechanisms improved visual-text alignment and

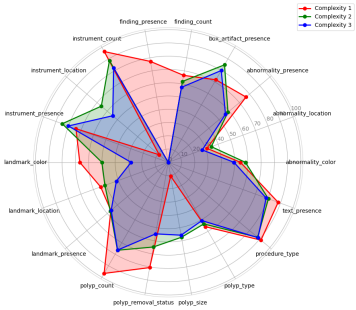
Table 1
Performance comparison of Subtask 1 Private and Public Test set results across different query complexity levels.

(a) Private Test Set

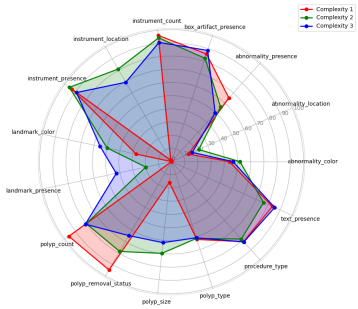
Level	R1	R2	RL	M	CHRF++	BLEU	F1
C1	0.39	0.23	0.36	0.54	51.99	0.10	0.90
C2	0.39	0.23	0.35	0.54	49.93	0.12	0.90
C3	0.47	0.31	0.38	0.63	55.88	0.20	0.91
Overall	**0.42**	**0.25**	**0.36**	**0.57**	**53.12**	**0.16**	**0.90**

(b) Public Test Set

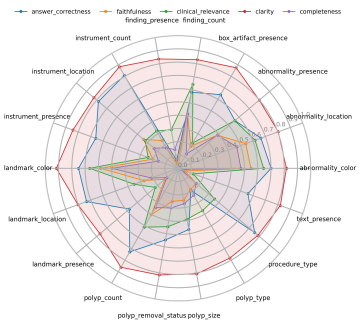
Level	R1	R2	RL	M	CHRF++	BLEU	F1
C1	0.38	0.23	0.36	0.54	52.59	0.10	0.89
C2	0.40	0.23	0.35	0.54	51.66	0.13	0.90
C3	0.47	0.29	0.38	0.62	55.67	0.19	0.90
Overall	**0.41**	**0.25**	**0.36**	**0.56**	**53.74**	**0.15**	**0.90**



(a) **Scores for public set across various question categories**



(b) **Scores for private set across various question categories**



(c) **Subtask 2 score distribution by question types**

Figure 2: Graph plots to visualize results of Subtask 1 and Subtask 2.

Table 2
Quantitative Results: Evaluation metrics per question type.(Subtask 2)

Question Type	Answer Correctness	Faithfulness	Clinical Relevance	Clarity	Completeness
abnormality_color	0.7012	0.5494	0.6446	0.8169	0.4735
abnormality_location	0.6647	0.5429	0.6179	0.8038	0.4712
abnormality_presence	0.5599	0.3880	0.5577	0.7993	0.3887
box_artifact_presence	0.6406	0.1924	0.2201	0.8750	0.1246
finding_count	0.5820	0.4140	0.6424	0.8314	0.4180
finding_presence	0.0597	0.0336	0.2955	0.8366	0.1433
instrument_count	0.8087	0.2413	0.3245	0.8837	0.1786
instrument_location	0.7830	0.3388	0.3234	0.8277	0.2351
instrument_presence	0.6601	0.2028	0.2399	0.8427	0.1573
landmark_color	0.7500	0.5833	0.6667	0.9167	0.5000
landmark_location	0.7331	0.2729	0.3517	0.8025	0.2144
landmark_presence	0.4788	0.1288	0.2246	0.7678	0.1102
polyp_count	0.7276	0.4041	0.5122	0.8612	0.3418
polyp_removal_status	0.5483	0.2534	0.4472	0.8165	0.3040
polyp_size	0.4684	0.2454	0.3897	0.8080	0.2707
polyp_type	0.1919	0.1848	0.3692	0.7813	0.2354
procedure_type	0.7543	0.1791	0.3594	0.7859	0.1782
text_presence	0.5603	0.0580	0.1223	0.8205	0.0263
Total	0.5942	0.2686	0.3842	0.8236	0.2455

explanation quality, while confidence estimation quantified epistemic certainty. For future work, we aim to extend this framework to larger vision-language models, incorporate advanced data augmentation and image enhancement techniques, refine reasoning over complex question types, and enhance model calibration.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Gautam, V. Thambawita, M. Riegler, et al., Medico 2025: Visual Question Answering for Gastrointestinal Imaging, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025. doi:10.48550/arXiv.2508.10869.

[2] S. Gautam, M. Riegler, P. Halvorsen, Kvasir-VQA-x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy, in: Data Engineering in Medical Imaging, Springer, 2025, pp. 53–63. doi:10.1007/978-3-032-08009-7_6.

[3] J. Lau, E. Lehman, P. Sharma, A dataset and exploration of models for understanding radiology reports, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2018, pp. 1–10. URL: https://aclanthology.org/D18-1001/.

[4] X. He, Y. Zhang, L. Mou, E. Xing, P. Xie, PathVQA: 30000+ Questions for Medical Visual Question Answering, arXiv preprint arXiv:2003.10286 (2020).

[5] 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: ACM Conferences, Association for Computing Machinery, New York, NY, USA, 2024, pp. 3–12. doi:10.1145/3689096.3689458.

[6] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, H. K. Johansen, M. A. Riegler, P. Halvorsen, HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, Scientific Data 7 (2020) 1–14. doi:10.1038/s41597-020-00622-y.

[7] S. A. Hicks, A. Storås, P. Halvorsen, T. de Lange, M. A. Riegler, V. Thambawita, Overview of ImageCLEFmedical 2023 – Medical Visual Question Answering for Gastrointestinal Tract, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[8] B. Ionescu, H. Müller, A.-M. Drăgulinescu, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, et al., Overview of the ImageCLEF 2024: Multimedia Retrieval in Medical Applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer, Cham, Switzerland, 2024, pp. 140–164. doi:10.1007/978-3-031-71908-0_7.

[9] S. A. Hicks, A. Storås, P. Halvorsen, M. A. Riegler, V. Thambawita, Overview of ImageCLEFmedical 2024 – Medical Visual Question Answering for Gastrointestinal Tract, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[10] B. Ionescu, H. Müller, D.-C. Stanciu, A. Idrissi-Yaghir, A. Radzhabov, A. G. S. de Herrera, A. Andrei, A. Storås, A. B. Abacha, B. Bracke, B. Lecouteux, B. Stein, C. Macaire, C. M. Friedrich, C. S. Schmidt, D. Fabre, D. Schwab, D. Dimitrov, E. Esperança-Rodier, G. Constantin, H. Becker, H. Damm, H. Schäfer, I. Rodkin, I. Koychev, J. Kiesel, J. Rückert, J. Malvehy, L.-D. undefinedtefan, L. Bloch, M. Potthast, M. Heinrich, M. A. Riegler, M. Dogariu, N. Codella, P. Halvorsen, P. Nakov, R. Brüngel, R. A. Novoa, R. J. Das, S. A. Hicks, S. Gautam, T. M. G. Pakull, V. Thambawita, V. Kovalev, W.-W. Yim, Z. Xie, Imageclef 2025: Multimedia retrieval innbsp;medical, social media andnbsp;content recommendation applications, in: Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V, Springer-Verlag, Berlin, Heidelberg, 2025, p. 398–406. URL: https://doi.org/10.1007/978-3-031-88720-8_60. doi:10.1007/978-3-031-88720-8_60.

[11] B. Ionescu, H. Müller, D.-C. Stanciu, A. Idrissi-Yaghir, A. Radzhabov, A. G. S. de Herrera, A. Andrei, A. Storås, A. B. Abacha, B. Bracke, B. Lecouteux, B. Stein, C. Macaire, C. M. Friedrich, C. S. Schmidt, D. Fabre, D. Schwab, D. Dimitrov, E. Esperança-Rodier, G. Constantin, H. Becker, H. Damm, H. Schäfer, I. Rodkin, I. Koychev, J. Kiesel, J. Rückert, J. Malvehy, L.-D. Ștefan, L. Bloch, M. Potthast, M. Heinrich, M. A. Riegler, M. Dogariu, N. Codella, P. H. P. Nakov, R. Brüngel, R. A. Novoa, R. J. Das, S. A. Hicks, S. Gautam, T. M. G. Pakull, V. Thambawita, V. Kovalev, W.-W. Yim, Z. Xie, Overview of ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications, booktitle = Experimental IR Meets Multilinguality, Multimodality, and Interaction,

Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.
[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Chen, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, in: Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR, 2021, pp. 11113–11125.
[13] R. Dutt, L. Ericsson, P. Sanchez, S. A. Tsaftaris, T. Hospedales, Parameter-Efficient Fine-Tuning for Medical Image Analysis: The Missed Opportunity, in: N. Burgos, C. Petitjean, M. Vakalopoulou, S. Christodoulidis, P. Coupe, H. Delingette, C. Lartizien, D. Mateus (Eds.), Proceedings of The 7th International Conference on Medical Imaging with Deep Learning (MIDL), volume 250 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 406–425. URL: https://proceedings.mlr.press/v250/dutt24a.html.
[14] K. Tewari, S. Pal, Advancing Vision and Language in GI Diagnosis: Florence2 for Question Answering and Stable Diffusion for Image Synthesis, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[15] K. Tewari, A. Verma, S. Pal, IReL, IIT(BHU) at MEDIQA-MAGIC 2025: Tackling Multimodal Dermatology with CLIPSeg-Based Segmentation and BERT-Swin Question Answering, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[16] M. Chaichuk, S. Gautam, S. Hicks, E. Tutubalina, Prompt to Polyp: Medical Text-Conditioned Image Synthesis with Diffusion Models, arXiv (2025). doi:10.48550/arXiv.2505.05573.
[17] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004, pp. 74–81.
[18] S. Banerjee, A. Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.
[19] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 311–318.
[20] M. Popović, chrF++: words helping character n-grams, in: Proceedings of the Second Conference on Machine Translation (WMT), Copenhagen, Denmark, 2017, pp. 612–618.