

Medico 2025: Visual Question Answering (with multimodal explanations) for Gastrointestinal Imaging

Sarath Chandran K.R.^{1,†}, Sivasriraman P^{2,†}, Vishnu Muruges V^{3,†} and Vishwajith L.K.^{4,†}

Department of Computer Science, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

Abstract

This work describes a PaliGemma 2-trained method for the Medico 2025 Visual Question Answering (VQA) challenge on GI imaging. Fine-tuning on the Kvasir-VQA-x1 dataset produces clinically sound answers to questions about colonoscopy. Interpretability is improved with a lightweight explanation module based on Grad-CAM visual maps, confidence estimation, and textual reasoning with structure. The system is shown to have robust generalization with little adaptation and enhancing clinical transparency and reliability.

1. Introduction

Correct interpretation of gastrointestinal (GI) endoscopy images is critical for early diagnosis and treatment. Visual Question Answering (VQA) offers an avenue for AI systems to answer clinical questions directly from medical images and enhance decision support in diagnosis. In this paper, we transform PaliGemma 2, a multimodal model which incorporates visual and text understanding, to the Medico 2025 challenge based on the Kvasir-VQA-x1 dataset. Our emphasis is in making model predictions more interpretable by ensuring generated responses are consistent with human-interpretable visual evidence. Model performance is measured based on common language evaluation metrics like BLEU, ROUGE, and METEOR, highlighting the potential and drawbacks of the model in medical VQA tasks.

2. Related Work

Vision-language model (VLM) advances over the past year have greatly enhanced computer-aided medical image interpretation. Models like CLIP [1], BLIP [2], and LLaVA [3] have shown robust multimodal reasoning capacities on clinical image-text tasks, such as visual question answering and diagnostic captioning. In endoscopy and radiology analysis, explainability methods like Grad-CAM [4] have been widely applied to improve interpretability by showing diagnostically important regions of interest in the images. Based on these, this work uses the PaliGemma 2 model to conduct clinically guided question answering and explainable interpretation of colonoscopy, connecting visual attention and structured clinical knowledge.

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


*Corresponding author.

[†]These authors contributed equally.

✉ sarathchandran@ssn.edu.in (S. C. K.R.); sivasriraman2370066@ssn.edu.in (S. P);
vishnumuruges2370054@ssn.edu.in (V. M. V); vishwajith2370033@ssn.edu.in (V. L.K)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

3. Task Overview and Dataset

The Medico 2025 Challenge [5] has two subtasks: Subtask 1 deals with clinically pertinent questions from GI endoscopy images, and Subtask 2 takes this further to multimodal explanations. Both employ the Kvasir-VQA-x1 dataset [6], which consists of approximately 6.5k original and 65k augmented images with 159k question-answer pairs involving multiple types (yes/no, color, location, numerical, choice-based) and difficulty levels. Every sample contains image, question, ground-truth answer, and complexity label. In this paper, we mainly tackle Subtask 1 with PaliGemma 2, and we have initial investigation on Grad-CAM-based visual interpretability for Subtask 2.

4. Methodology

4.1. Subtask 1: Visual Question Answering (VQA)

The solution proposed for Subtask 1 employs **PaliGemma 2**, an open-weight, unified **Vision-Language Model (VLM)**, to generate contextually accurate, free-form answers from endoscopy images and clinical questions.

4.1.1. Model Architecture

PaliGemma 2 adopts a modular encoder-decoder Transformer structure, enabling strong cross-modal reasoning:

Vision Encoder: We utilize the *SigLIP (So400m/14)* Vision Transformer to extract robust spatial and textural features from the input endoscopy image (e.g., at 448×448 resolution), which is essential for detailed mucosal and lesion analysis.

Text Language Encoder/Decoder: Derived from the **Gemma** family of LLMs, this module interprets the clinical question and performs **autoregressive generation** of the final answer.

Cross-Modal Integration: Visual tokens are **linearly projected** and **concatenated** with text embeddings. The decoder’s **cross-attention layers** dynamically ground the generated answer tokens to the relevant visual regions, ensuring clinical fidelity.

4.1.2. Fine-Tuning Strategy

A comprehensive fine-tuning pipeline is employed to specialize the VLM for gastrointestinal (GI) diagnostics:

Dataset Adaptation: The model is fine-tuned end-to-end on the *Kvasir-VQA-x1* dataset, using a prompt format:

Question: [query] Answer: [answer]

Optimization Objective: **Cross-entropy loss** is minimized over the predicted answer tokens, ensuring both lexical and semantic correctness.

Domain Alignment: Full fine-tuning internalizes GI-specific terminology and pathology patterns, effectively bridging the general-domain performance gap and tailoring the model to the diagnostic needs of the Medico 2025 VQA challenge.

5. Results and Analysis

This section reports the performance of the fine tuned PaliGemma 2 on **Subtask 1 (Visual Question Answering)** using both the *public* and *private* test sets. Evaluation employs standard VQA metrics, including ROUGE-1/2/L, METEOR, CHRF++, BLEU, and BERTScore (Precision/Recall/F1).

5.1. Subtask 1: VQA Results

5.1.1. Private Test Set

Results for the private test set are shown in Table 1 and visualized with the radar plot in Figure 1b.

Table 1
Subtask 1 VQA Results on Private Test Set

Level	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CHRF++	BLEU	BERTScore (P/R/F1)
Complexity 1	0.6972	0.6062	0.6852	0.6502	64.15	0.4258	0.9480 / 0.9486 / 0.9105
Complexity 2	0.6298	0.4171	0.5949	0.6126	57.89	0.3216	0.9381 / 0.9397 / 0.9388
Complexity 3	0.6917	0.5125	0.6459	0.6634	61.85	0.4510	0.9494 / 0.9448 / 0.9470
Overall	0.6732	0.4788	0.6424	0.6418	60.88	0.4225	0.9451 / 0.9444 / 0.9447

Table 2
Subtask 1 VQA Results on Public Test Set

Level	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CHRF++	BLEU	BERTScore (P/R/F1)
Complexity 1	0.6987	0.5229	0.6912	0.6680	65.59	0.4224	0.9517 / 0.9528 / 0.9521
Complexity 2	0.6519	0.4467	0.6244	0.6293	61.01	0.3652	0.9431 / 0.9429 / 0.9429
Complexity 3	0.6996	0.5123	0.6510	0.6624	61.98	0.4473	0.9493 / 0.9426 / 0.9458
Overall	0.6837	0.4948	0.6562	0.6534	62.26	0.4273	0.9481 / 0.9462 / 0.9470

5.1.2. Public Test Set

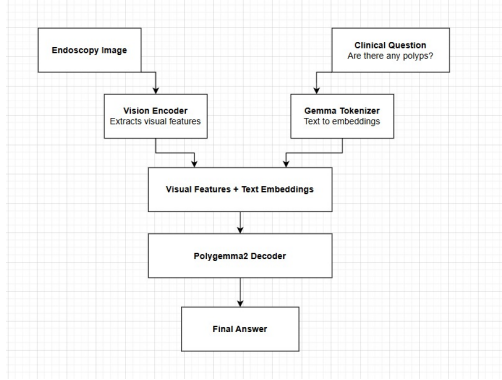
Results for the public test set are reported in Table 2

5.2. Analysis

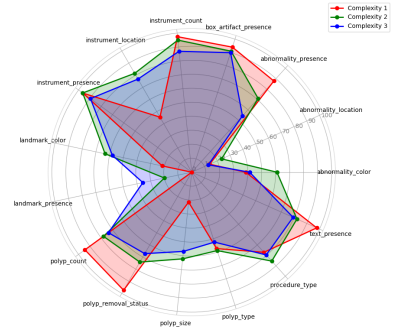
Strong semantic and lexical agreement between predictions and references are shown by High ROUGE-L (0.65), METEOR (0.64–0.65), and BERTScore (F1 0.94–0.95). Fair n-gram precision is shown by moderate BLEU (0.42) and CHRF++ (61–62). Complexity 1 has the highest values, with minor drops at increased complexities, demonstrating better performance on easier visual-question pairs while retaining strong generalization across sets.

6. Discussion and Comparison

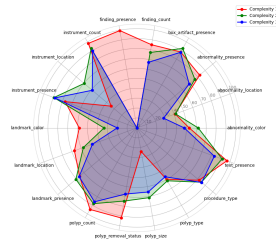
The PaliGemma 2 VQA model achieved strong performance on the Public Test set, significantly surpassing general baselines. Quantitatively, the model delivered high semantic accuracy with ROUGE-1 at approximately 0.68 and an excellent BERTScore F1 of approximately 0.947,



(a) Model architecture of PaliGemma 2 for VQA.



(b) Radar plot on the private test set.



(c) Radar plot on the public test set.

Figure 1: Visualization of Subtask 1 results. (a) PaliGemma 2 architecture overview. (b–c) Radar plots comparing evaluation metrics on private and public test sets.

confirming its fidelity to clinical ground truth. A BLEU score of approximately 0.427 also demonstrated good lexical precision. Notably, the model’s performance on Complexity 3 (Multi-step Reasoning) remained robust, with METEOR scores reaching approximately 0.66, validating the strong reasoning capabilities of the Gemma decoder over combined visual features. This tailored VLM provides a competitive benchmark for VQA and establishes a solid foundation for future work on interpretability and clinical adoption.

7. Conclusion and Future Work

This research demonstrated the effectiveness of integrating multimodal big language models, especially PaliGemma 2, during the interpretation of colonoscopy images through visual question answering and Grad-CAM–based explanations. Future works will focus on generating outputs in a clinically formatted manner to enhance interpretability and clinical significance and fine-tuning in a domain-specific context with increased datasets.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

- [1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, Proceedings of the 38th International Conference on Machine Learning (2021).
- [2] J. Li, D. Li, C. Xiong, S. C. Hoi, Blip: Bootstrapped language-image pre-training for unified vision-language understanding and generation, in: Proceedings of the 39th International Conference on Machine Learning, 2022.
- [3] H. Liu, C. Li, Q. Wu, Y. J. Lee, Llava: Visual instruction tuning, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).
- [4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
- [5] S. Gautam, V. Thambawita, M. Riegler, et al., Medico 2025: Visual Question Answering for Gastrointestinal Imaging, arXiv (2025). doi:10.48550/arXiv.2508.10869. arXiv:2508.10869.
- [6] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-VQA-x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy, arXiv (2025). doi:10.48550/arXiv.2506.09958. arXiv:2506.09958.