Exploring Vision-Language Models for Medical VQA on Gastrointestinal Images: A LoRA Fine-Tuning Study

Vishal Subramoniam[1], Rajalakshmi Sivanaiah[2] and Angel Deborah S[3]

SSN College of Engineering, India

Abstract

This work addresses Task 1 (Subtask 1) of the 2025 MediaEval Medico Challenge, focusing on the development of explainable AI models for Medical Visual Question Answering (VQA) in gastrointestinal (GI) imaging. We explored and tested various multimodal models, including Florence (Microsoft), Salesforce BLIP, and Google PaliGemma, on the Kvasir-VQA-x1 dataset. After evaluating their baseline performance, we fine-tuned these models on a subset of cases, and finally used the PaliGemma model using Low-Rank Adaptation (LoRA) on the full dataset of ∼159,000 samples. The fine-tuned model demonstrated competitive performance across multiple VQA categories, advancing the integration of multimodal data and improving the explainability of AI-driven solutions in GI imaging.

https://huggingface.co/vsl366/KvasirMedVQA

1. Introduction

Gastrointestinal (GI) disorders are a leading cause of global morbidity, with endoscopy serving as the primary modality for diagnosis and treatment guidance [1]. Despite its clinical importance, manual interpretation of endoscopic images remains labor-intensive and prone to inter-observer variability, motivating the need for automated, explainable AI systems to assist clinicians in visual analysis.

Visual Question Answering (VQA) represents a crucial step toward multimodal clinical reasoning, as it unifies image understanding and language comprehension to provide interpretable, text-based responses to image-grounded questions. Such models enhance transparency and trust in medical AI, particularly in domains requiring both diagnostic accuracy and semantic interpretability. Within the **Medico 2025 VQA Challenge** [2], **Subtask 1** focuses on generating clinically relevant answers to questions about GI endoscopic images. In this work, we explore the application of recent **Vision-Language Models** (VLMs), namely **Florence** [3], **BLIP** [4], and **PaliGemma** [5] on the **Kvasir-VQA** [6] dataset comprising **159k** question-answer pairs from 6.5k GI images across diverse anatomical and pathological categories.

To adapt these large pretrained models to the medical domain efficiently, we employ **Low-Rank Adaptation** (LoRA) [7] within a Parameter-Efficient Fine-Tuning (PEFT) framework, enabling targeted domain adaptation without full-model retraining. Through proper evaluation, we examine how architecture, data scale, and fine-tuning strategies affect medical VQA performance and reliability. Our study establishes strong baselines for GI-based multimodal reasoning and provides insights into building parameter-efficient, explainable systems.

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online
vishal2310293@ssn.edu.in (V. Subramoniam); rajalakshmis@ssn.edu.in (R. Sivanaiah); angeldeborahs@ssn.edu.in (A. D. S)
0009-0002-2822-8472 (V. Subramoniam); 0000-0003-4658-5465 (R. Sivanaiah); 0000-0003-1197-2852 (A. D. S)

2. Related Work

Research in medical Visual Question Answering has evolved significantly with large-scale multimodal datasets and vision–language models for clinical imaging. Early work such as VQA-RAD [8] pioneered radiology-based visual reasoning, establishing foundational benchmarks for medical VQA. Subsequently, gastrointestinal imaging gained attention with the introduction of Kvasir-VQA, and the recently released Kvasir-VQA-x1 [9], which provides 159,549 question–answer pairs with complexity annotations for robust multimodal learning. Recent studies have integrated general-purpose Vision–Language Models into the medical domain. Florence-2, introduced by Microsoft, presents a unified, prompt-based representation for various computer vision and vision-language tasks. BLIP pioneered bootstrapping language-image pre-training for unified vision-language understanding and generation, achieving state-of-the-art results on image-text retrieval, captioning, and VQA tasks. PaliGemma, developed by Google, offers a versatile 3B parameter model combining SigLIP vision encoding with Gemma language modeling, demonstrating strong transfer learning capabilities. Methods such as Prompt-to-Polyp [10] explored diffusion-based text–image synthesis for domain adaptation, while ImageCLEF medical tasks [11, 12] have benchmarked these models for diagnostic reasoning. PEFT approaches, particularly LoRA, have demonstrated competitive accuracy with significantly reduced computational requirements—reducing trainable parameters by up to 10,000 times compared to full fine-tuning. Moreover, TrustVLM [13] emphasized hallucination control and uncertainty estimation in clinical VQA systems, highlighting the need for transparent and reliable responses in safety-critical medical applications.

3. Task Overview and Dataset

We participated in **Subtask 1 (Visual Question Answering, VQA)** of the *MediaEval Medico 2025* challenge, which aims to advance the development of trustworthy and interpretable multimodal systems for gastrointestinal (GI) image understanding. The subtask focuses on generating clinically meaningful natural language answers to visual questions related to GI endoscopic images, thereby supporting diagnostic and educational use cases in medical AI. Our experiments utilized the **Kvasir-VQA-x1** dataset, comprising **159,549 question–answer pairs** derived from **6,500 original GI images**. The dataset extends the original Kvasir-VQA and HyperKvasir by introducing additional question complexity scores, facilitating controlled training and benchmarking across varying difficulty levels. The questions span clinically relevant domains, including anatomical structures, pathological findings, therapeutic tools, and procedural steps, reflecting real-world diagnostic diversity. We adopted the official baseline notebook and validation script provided by the challenge organizers as our implementation framework. Initial experiments were conducted on a 10k subset of QA pairs, split 80:10 for training and validation, before scaling to the full dataset. Images and corresponding question–answer pairs were preprocessed following model-specific requirements—resized to 224×224 for Florence and BLIP, and 256×256 for PaliGemma, alongside standardized text tokenization.

4. Methodology

We develop an multimodal VQA pipeline to generate clinically accurate answers from gastrointestinal (GI) endoscopic images. Our approach leverages three state-of-the-art vision–language models: **Microsoft Florence-2-large** (Florence-L), **Salesforce BLIP-VQA-CapFILT-large** (BLIP-L), and **Google PaliGemma-3B-pt-224** (PaliGemma). Each model follows a multi-

modal architecture: a frozen vision encoder extracts high-dimensional image embeddings, a transformer-based text encoder processes the input question, and a multimodal fusion mechanism integrates both modalities for autoregressive answer generation. Florence-L and BLIP-L use embedding concatenation and linear projection, while PaliGemma employs cross-attention between the decoder and visual embeddings, enabling selective focus on diagnostically relevant regions. This design enhances grounding and reduces hallucinations in medical responses. To improve visual grounding and avoid overfitting to textual priors, random image–question mismatches were introduced during training. We applied parameter-efficient **LoRA fine-tuning** to the attention and projection layers, keeping the vision backbone frozen to prevent catastrophic forgetting. Gradient checkpointing and 4-bit NF4 quantization allowed efficient training with extremely limited resources (Free Tier Colab T4 GPU, 16 GB Ram), thereby sacrificing time. For preliminary experiments, a subset of **10,000 QA pairs** from Kvasir-VQA-x1 was used, consisting of 1000 random images of varying types, with 10 questions each, split 80:20 for training and validation. Various but a limited set of hyperparameters, as shown in Table 1 enabled rapid prototyping and comparison. These parameters were chosen as they were the standard ones. The final configuration scaled PaliGemma (due to its better output) fine-tuning to the full dataset.

Table 1
LoRA fine-tuning hyperparameters for subset and full-dataset training (large variants).

Model	Subset	LoRA (r,α)	LR $\times 10^{-5}$	Ep	Batch	Grad Acc	Sch	Wt Decay
Florence-L	10k	8,16	5	2	4	1	Cos	0.01
BLIP-L	10k	12,24	3	2	3	1	Lin	0.01
PaliGemma	10k	16,32	2	1	4	4	Lin	0.01
PaliGemma	159k	16,32	2	1	4	4	Lin	0.01

5. Results and Evaluation

The comparison of the three mentioned model on a subset of rows, summarized in Table 2, shows PaliGemma outperforming the other two models across all text-generation metrics. The improvement is attributed to its cross-attention fusion and larger multimodal capacity. All subset experiments used LoRA-based fine-tuning under identical conditions for fairness.

Table 2
Subset (10k) results across LoRA Fine-tuned models using BLEU, ROUGE-1, and ROUGE-L metrics.

Model	BLEU	ROUGE-1	ROUGE-L
Florence-L	0.512	0.785	0.770
BLIP-L	0.529	0.798	0.781
PaliGemma-3B	**0.553**	**0.818**	**0.795**

The final fine-tuned PaliGemma model achieved strong consistency across all complexity levels on both the **Full** and **Private** evaluation sets. As shown in Table 3, the model maintained balanced performance across easy (Level 1), moderate (Level 2), and complex (Level 3) questions. The highest overall performance was achieved on the Full set with **ROUGE-1 = 0.724**, **BLEU = 0.494**, and **BERTScore F1 = 0.9548**. Results on the Private set remained comparable, indicating strong generalization to unseen data. Radar plots provided by the organizers (Figure 1) visualize absolute performance across question categories and complexity levels. The PaliGemma model demonstrated consistent alignment across clinically critical categories, suggesting strong visual grounding and robustness to linguistic variation.

Table 3
Performance across complexity levels for Full and Private sets

Level	ROUGE-1		ROUGE-L		METEOR		BLEU		BERTScore	
	Full	Private	Full	Private	Full	Private	Full	Private	Full	Private
Level 1	0.73	0.70	0.72	0.69	0.69	0.65	0.50	0.46	0.96	0.95
Level 2	0.69	0.68	0.66	0.65	0.67	0.65	0.42	0.39	0.95	0.95
Level 3	0.75	0.74	0.71	0.69	0.75	0.73	0.53	0.52	0.96	0.96

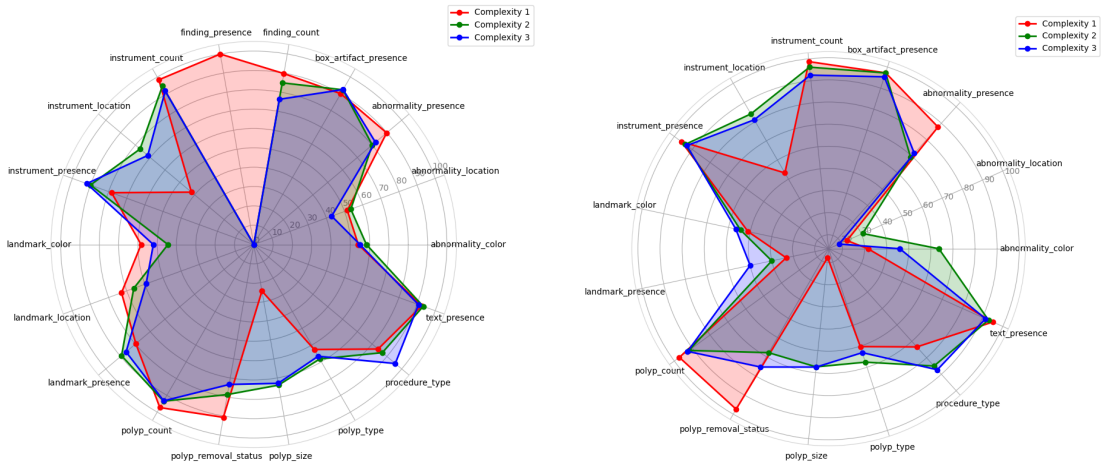

Figure 1: Radar plots for Full (L) and Private (R) evaluation sets showing per-category performances.

6. Discussion and Comparison with Literature

Our fine-tuned **PaliGemma-3B** model slightly outperformed prior baselines on the Kvasir-VQA-x1 dataset, improving ROUGE-L by approximately 1.2% and BLEU by about 3.8% over the baseline. Compared with Florence-L and BLIP-L, PaliGemma's cross-attention fusion achieved better visual grounding and fewer hallucinations, aligning with recent findings that decoder-based attention enhances medical VQA reliability. The main limitations were small absolute gains near the dataset's performance ceiling and occasional hallucinations in knowledge-heavy questions. Moreover, automatic text metrics often penalized clinically correct but lexically divergent answers. Finally, due to limited resources, Ablation studies were not done.

7. Conclusion and Future Work

We have tested various different models with various parameters on the dataset. Future directions include extending the model to multimodal explanations, integrating knowledge retrieval for complex reasoning, and adding uncertainty calibration. Robustness testing under image perturbations and human-in-the-loop validation are planned to ensure trustworthy VQA systems.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 in order to correct grammatical and spelling errors. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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, et al., 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.

[2] 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.

[3] 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024, pp. 4818–4829. doi:10.1109/CVPR52733.2024.00463.

[4] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, in: Proceedings of the 39th International Conference on Machine Learning (ICML), volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 12888–12900.

[5] L. Beyer, A. Steiner, R. Bachmann, X. Deng, O. J. Henaff, M. Minderer, A. Arnab, M. Dehghani, F. Pavetic, J. Shlens, et al., PaliGemma: A versatile 3B VLM for transfer, arXiv preprint arXiv:2407.07726 (2024). doi:10.48550/arXiv.2407.07726.

[6] 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.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, in: International Conference on Learning Representations (ICLR), 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[8] A. B. Abacha, S. A. Hasan, V. Datla, J. Liu, D. Demner-Fushman, H. Muller, VQA-RAD: Visual Question Answering in Radiology, Medical Image Analysis 68 (2021) 101871. doi:10.1016/j.media.2020.101871.

[9] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-VQA-x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy, arXiv preprint arXiv:2506.09958 (2025). doi:10.48550/arXiv.2506.09958.

[10] M. Chaichuk, S. Gautam, S. Hicks, E. Tutubalina, Prompt to Polyp: Medical Text-Conditioned Image Synthesis with Diffusion Models, arXiv preprint arXiv:2505.05573 (2025). doi:10.48550/arXiv.2505.05573.

[11] 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.

[12] B. Ionescu, H. Müller, D.-C. Stanciu, A. Idrissi-Yaghir, A. Radzhabov, et al., ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications, in: Guide Proceedings, Springer-Verlag, Berlin, Germany, 2025, pp. 398–406. doi:10.1007/978-3-031-88720-8_60.

[13] Z. Zhang, X. Liu, Y. Wang, Q. Li, H. Chen, Evaluating Trust and Hallucination in Medical Vision–Language Models, Nature Machine Intelligence 6 (2024) 1067–1079. doi:10.1038/s42256-024-00958-9.