From Answers to Explanations: Self-Probing Efficiently Fine-Tuned Vision–Language Models for Medical VQA at Medico 2025

Sujata Gaihre[1], Amir Thapa Magar[1]

[1]NCIT, Kathmandu, Nepal

Abstract

We present our submission to the **MediaEval Medico 2025 Challenge**, addressing **Subtask 1 (Visual Question Answering)** and **Subtask 2 (Multimodal Explanations)** in gastrointestinal (GI) imaging. Our system fine-tunes the general-purpose vision–language model `google/paligemma-3b-pt-224` on the `Kvasir-VQA-x1` dataset using Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) and 4-bit quantization. A compact augmentation pipeline introducing geometric and color-space variations enhances robustness and generalization. For Subtask 2, we generate textual explanations through a prompt-based approach and evaluate interpretability across correctness, faithfulness, relevance, clarity, and completeness. On the private test set, our approach achieves a ROUGE-1 score of **0.723** and a METEOR score of **0.70** for Subtask 1, along with strong performance across the interpretability metrics in Subtask 2. Code and model checkpoints are available at: `github.com/sujata-gaihre/MediaEval-Medico-2025-paligemma`.

1. Introduction and Related Work

Accurate interpretation of endoscopic imagery is crucial for diagnosing gastrointestinal (GI) diseases, yet manual review remains time-consuming and subjective [1]. The **Medico 2025 Challenge** [2] promotes multimodal AI systems that enhance both diagnostic accuracy and interpretability through two subtasks: **Visual Question Answering (Subtask 1)** and **Multimodal Explanations (Subtask 2)**.

Medical VQA datasets such as VQA-RAD [3], PathVQA [4], and Kvasir-VQA [5, 6] have enabled the study of multimodal reasoning in clinical imaging [7]. Challenges like ImageCLEF VQA-Med [8, 9] further underscore the dual need for accuracy and explainability in clinical AI. Meanwhile, large-scale multimodal models such as CLIP [10], BLIP [11], and Flamingo [12]—have demonstrated the transferability of vision–language reasoning through joint pre-training.

Building upon these foundations, our approach leverages the `google/paligemma-3b-pt-224` model, which integrates Gemma's language backbone with visual encoders [13, 14]. We fine-tune this model on the GI-specific `Kvasir-VQA-x1` dataset using Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) [15, 16], along with lightweight image transformations to improve robustness under constrained computational settings [17]. The resulting unified framework effectively addresses both Medico subtasks—producing accurate answers and interpretable text-based explanations for endoscopic imagery.

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

sujatag391@gmail.com (S. Gaihre); amirthapa2019@gmail.com (A. T. Magar)

0009-0001-8739-9258 (S. Gaihre)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Methodology

Our framework addresses both Subtask 1 and Subtask 2 by fine-tuning a single, powerful vision-language model (VLM), `google/paligemma-3b-pt-224`. The core of our contribution lies in a custom data augmentation strategy to enhance model robustness and a novel two-stage, self-probing pipeline to generate comprehensive explanations.

2.1. Data Preprocessing, Augmentation, and Tuning

We utilized the Medico 2025 challenge dataset, Kvasir-VQA-x1, a large-scale benchmark containing 159,549 complex question-answer pairs derived from 6,500 GI images. As per our processing script, we used the entire pre-defined "train" and "test" splits without creating a separate validation set. To improve generalization, we implemented a data augmentation pipeline for the **training split** using `albumentations`, including: **Geometric Transformations** (an `Affine` transform with **rotation** [±10°], **translation** [up to 10%], and **scaling** [90–110%]) and **Color Space Transformations** (`ColorJitter` for brightness, contrast, saturation, and hue).

We selected `google/paligemma-3b-pt-224` and employed MS-Swift for efficient training on a single TESLA T4 GPU. We utilized **QLoRA** (4-bit NF4 quantization + LoRA), fine-tuning for **2 epochs** with the **AdamW** optimizer. Key hyperparameters were a learning rate of 2×10^{-5}, effective batch size of **16**, LoRA rank (r) of **16**, and alpha (α) of **32**.

2.2. Task-Specific Pipelines

Subtask 1: Visual Question Answering. For the primary VQA task, the fine-tuned model autoregressively generates a direct textual answer when provided with an image-question pair.

Subtask 2: Explanation Generation. We designed a **two-stage, self-probing pipeline** that leverages both our VLM and an LLM for synthesis.

1.) **Initial Answering**: Our fine-tuned VLM generates a primary answer to the user's original question.

2.) **Self-Probing**: We created a comprehensive mapping (detailed in our code) that links 19 `question_class` categories to 10 pre-defined, targeted follow-up questions each. This mapping forces the model to probe for specific, granular details, as exemplified in Table 1.

3.) **Explanation Synthesis**: The primary answer and probing question-answer pairs are fed into OpenAI's `gpt-4o-mini` model, which is prompted to synthesize them into a single, coherent textual explanation.

Table 1

Simplified example from our custom mapping for the self-probing mechanism. Each question class triggers a set of targeted, pre-defined questions.

Question Class	Example Pre-defined Probing Questions
`abnormality_location`	- What type of abnormality is at this location? - What is the size of the abnormality? - What is the color of the abnormality?
`instrument_presence`	- Where in the image is the instrument located? - How many instruments are present? - What type of instrument is visible?
`polyp_size`	- Where is the polyp located in the image? - What type of polyp is it? - Is the polyp sessile or pedunculated?

3. Results and Evaluation

The answers were evaluated using BLEU, ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), METEOR, CHRF++, and BERTScore. The performance was measured on the `Kvasir-VQA-x1` [5] test set and a private challenge set to assess generalization. A comprehensive summary of our results is presented in Table 2.

Table 2
Performance comparison across datasets and evaluation settings. BLEU, ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), METEOR, CHRF++, and BERTScore-F1 metrics are reported. Complexity levels (*Level 1–3*) are shown only for the *Full Kvasir-VQA-x1 Test Set* and *Private Challenge Set*, while the *Public Leaderboard* section reports aggregate results without complexity differentiation.

Setting	BLEU	R1	R2	RL	METEOR	CHRF++	BERT-F1
Public Leaderboard (1500 samples)							
Normal	0.4582	0.6955	0.5058	0.6677	0.6713	–	–
Augmented	0.4873	0.7184	0.5337	0.6899	0.6983	–	–
Kvasir-VQA-x1 Test Set							
Level 1	0.4958	0.7272	0.5630	0.7201	0.6878	68.62	0.9569
Level 2	0.4208	0.6905	0.4940	0.6641	0.6645	65.52	0.9493
Level 3	0.5332	0.7531	0.5736	0.7059	0.7513	70.71	0.9581
Overall	**0.4936**	**0.7238**	**0.5438**	**0.6970**	**0.7009**	**68.57**	**0.9548**
Private Challenge Set (500 unseen images, 5,368 QA pairs)							
Level 1	0.4614	0.7004	0.5183	0.6875	0.6452	65.43	0.9484
Level 2	0.3878	0.6782	0.4764	0.6470	0.6542	62.30	0.9460
Level 3	0.5169	0.7357	0.5624	0.6930	0.7344	67.95	0.9576
Overall	**0.4708**	**0.7046**	**0.5182**	**0.6756**	**0.6762**	**65.58**	**0.9505**

3.1. Subtask 1: Visual Question Answering

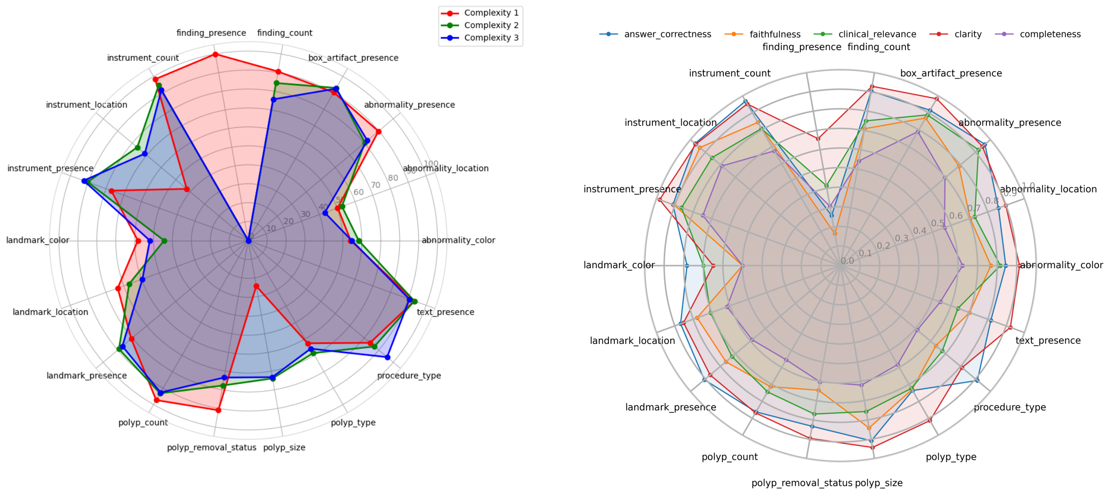
As shown in Table 2, our `albumentations` pipeline provided a significant performance boost. On the public leaderboard subset, data augmentation improved the ROUGE-1 score from 0.6955 to 0.7184 and the METEOR score from 0.6713 to 0.6983, confirming its value in enhancing model robustness.

On the `Kvasir-VQA-x1` test set, our final model achieved a strong overall ROUGE-1 score of 0.7238 and a BERTScore-F1 of 0.9548. Notably, the model excelled on the most complex (Level 3) questions, achieving the highest scores in this category (e.g., 0.7531 ROUGE-1), which underscores its advanced reasoning capabilities.

To validate generalization, the model was evaluated on a private challenge set of 500 previously unseen images. It maintained high performance, securing an overall ROUGE-1 of 0.7046 and a METEOR score of 0.6762. This consistency across datasets confirms that our approach generalizes effectively to new clinical data.

3.2. Subtask 2: Multimodal Explanation

Subtask 2 was evaluated using Qwen3-30B-A3B [18]as an automated adjudicator across five dimensions: Correctness, Faithfulness, Clinical Relevance, Clarity, and Completeness. The model scored highest in Correctness (0.8567) and Clarity (0.9044), as detailed in Table 3. Figure 1(b) visually breaks down these scores, highlighting strong performance on instrument-related queries but a significant weakness on “finding_presence” questions, which scored only 0.2590 in correctness.



(a) Subtask 1 performance by question type.

(b) Subtask 2 explanation scores by question type.

Figure 1: Qualitative performance analysis. (a) visualizes the absolute performance scores (e.g., ROUGE-1) of the three complexity levels across various question categories on the full Kvasir-VQA-x1 test set, highlighting strengths in instrument detection and weaknesses in counting tasks. (b) shows the breakdown of Subtask 2 explanation metrics, revealing low scores for finding_presence.

Table 3: Subtask 2 mean scores across question types.

Question Type	Correctness	Faithfulness	Relevance	Clarity	Completeness
abnormality_presence	0.9627	0.7894	0.9197	0.9458	0.6965
instrument_presence	0.9101	0.9028	0.8635	0.9826	0.7478
instrument_location	0.9691	0.9367	0.8564	0.9638	0.7904
finding_presence	0.2590	0.1649	0.4142	0.6575	0.3067
Overall	**0.8567**	**0.7421**	**0.7556**	**0.9044**	**0.6135**

4. Discussion and Outlook

Our fine-tuned PaliGemma model, using PEFT and data augmentation, proved effective for GI-related VQA, showing strong generalization on unseen data. However, the evaluation revealed a critical weakness in discerning the presence versus absence of findings (finding presence). While explanations were clear, they lacked sufficient faithfulness and completeness, indicating a reasoning disconnect. Future work will focus on improving presence/absence detection and exploring advanced explainability techniques. Deeper collaboration with clinicians is essential to bridge the gap to real-world clinical utility.

Acknowledgments

We thank the organizers of the MediaEval 2025 Medico Challenge and SimulaMet for their efforts in preparing the competition and the Kvasir-VQA-x1 dataset.

Declaration on Generative AI

During the preparation of this work, the authors utilized generative AI tools. Specifically, Grammarly was used for proofreading and improving grammar, while ChatGPT was used to enhance the clarity and readability of sentences. The authors reviewed and edited all content to ensure its accuracy and take full responsibility for the final manuscript.

References

[1] S. Lobanovs, J. Aleksejeva, A. K. Rūtiņa, E. Krustiņš, J. Čižovs, D. Bļizņuks, Machine learning in gastrointestinal endoscopy: challenges and opportunities, BMJ Open Gastroenterology 12 (2025).

[2] S. Gautam, V. Thambawita, M. Riegler, P. Halvorsen, S. Hicks, Medico 2025: Visual Question Answering for Gastrointestinal Imaging, ArXiv e-prints (2025). doi:10.48550/arXiv.2508.10869. arXiv:2508.10869.

[3] Y. Bazi, M. M. A. Rahhal, L. Bashmal, M. Zuair, Vision–language model for visual question answering in medical imagery, Bioengineering 10 (2023) 380.

[4] U. Naseem, M. Khushi, J. Kim, Vision-language transformer for interpretable pathology visual question answering, IEEE journal of biomedical and health informatics 27 (2022) 1681–1690.

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

[7] S. Gautam, M. A. Riegler, P. Halvorsen, Point, detect, count: Multi-task medical image understanding with instruction-tuned vision-language models, in: 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS), 2025, pp. 415–422. doi:10.1109/CBMS65348.2025.00090.

[8] 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.

[9] 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).

[10] S. Eslami, C. Meinel, G. De Melo, Pubmedclip: How much does clip benefit visual question answering in the medical domain?, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 1181–1193.

[11] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International conference on machine learning, PMLR, 2023, pp. 19730–19742.

[12] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, et al., Openflamingo: An open-source framework for training large autoregressive vision-language models, arXiv preprint arXiv:2308.01390 (2023).

[13] 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).

[14] A. Steiner, A. S. Pinto, M. Tschannen, et al., PaliGemma 2: A Family of Versatile VLMs for Transfer, ArXiv e-prints (2024). doi:10.48550/arXiv.2412.03555. arXiv:2412.03555.

[15] Z. Han, C. Gao, J. Liu, J. Zhang, S. Q. Zhang, Parameter-efficient fine-tuning for large models: A comprehensive survey, arXiv preprint arXiv:2403.14608 (2024).

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., ICLR 1 (2022) 3.

[17] A. Buslaev, A. Parinov, E. Khvedchenya, V. I. Iglovikov, A. A. Kalinin, Albumentations: fast and flexible image augmentations, ArXiv e-prints (2018). arXiv:1809.06839.

[18] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).