LoRA-Enhanced PaliGemma for Efficient Visual Question Answering in Gastrointestinal Imaging

Prabhash Kumar Jha[1], Firoj Paudel[2] and Debesh Jha[3]

[1]Dept. of Computer & Electronics Engineering, Kathford International College of Engineering and Management, Nepal
[2]Department of Computer Science, Madan Bhandari Memorial College, Nepal
[3]Department of Computer Science, University of South Dakota, USA

Abstract

This paper presents our approach to the MediaEval Medico 2025 Visual Question Answering (VQA) challenge for gastrointestinal (GI) imaging, using the Kvasir-VQA-x1 dataset. We fine-tune the PaliGemma-3B model using Low-Rank Adaptation (LoRA) to achieve efficient and high-performing VQA on 159,549 QA pairs from 6,500 GI endoscopic images. Our method leverages 4-bit quantization and a cosine learning rate scheduler to optimize performance while minimizing computational cost. We achieved BLEU 0.3973, ROUGE-1 0.6465, ROUGE-2 0.4436, ROUGE-L 0.6136 and METEOR 0.6119 on test data, adhering to the challenge's evaluation criteria. Our results demonstrate the effectiveness of LoRA for medical VQA, balancing accuracy and resource efficiency. The full implementation code is available at https://github.com/prabhashj07/MedEval_Medico_2025.

1. Introduction

Endoscopic image analysis is important for the detection of early gastrointestinal (GI) diseases, including colorectal cancer [1]. The MediaEval Medico 2025 challenge utilizes the Kvasir-VQA-x1 dataset [2] to make meaningful advances in trustworthy AI for medicine using Visual Question Answering (VQA). Subtask 1 focuses on developing accurate VQA models for diverse question types across 159,549 QA pairs from 6,500 GI images. We use a 4-bit quantized train split of the Kvasir-VQA-x1 dataset of 3,000 samples to fine-tune the model PaliGemma-3B using LoRA, to maximize performance on Subtask 1. Further, we investigate explanations of the model output using both text and visual formats at the end of the challenge. This paper reports the results from using commonly used evaluation metrics (BLEU, ROUGE, and METEOR) to assess our approach to developing interpretable, efficient AI for GI diagnosis.

2. Related work

Medical Visual Question Answering (VQA) facilitates clinical decision-making in gastrointestinal (GI) imaging. While early datasets like VQA-RAD [3] and PathVQA [4] focused on radiology and pathology with limited question diversity. The Kvasir-VQA dataset [5], built on HyperKvasir [6], introduced 6,500 GI endoscopy images with text-image annotations. Further, the Kvasir-VQA-x1 dataset [2] improved upon Kvasir-VQA by adding 159,549 unique QA pairs, a varied set of question types (e.g., Yes/No and location-type), and questions of an increased complexity level to enable new modes of multimodal reasoning.

Challenges like ImageCLEF 2024 and 2025 [7, 8] drove progress in medical multimedia and VQA, emphasizing interpretable AI, a focus continued in MediaEval Medico 2025 [1]. Efficient

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online
prabhashj07@gmail.com (P. K. Jha); firoj7902@mbmcsit.edu.np (F. Paudel); debesh.jha@usd.edu (D. Jha)
https://debeshjha.com/ (D. Jha)

fine-tuning methods like LoRA [9] and PEFT enable adaptation of models like PaliGemma-3B with low resource costs. Recent work like Prompt-to-Polyp [10] explores GI-specific vision-language tasks; importantly, the goal of the work is to understand image synthesis as opposed to VQA.

3. Methodology

To address Subtask 1 of the MediaEval Medico 2025 challenge on Visual Question Answering (VQA) for gastrointestinal (GI) imaging, we developed a robust pipeline leveraging the PaliGemma-3B vision-language model (`google/paligemma-3b-pt-224`), fine-tuned with Low-Rank Adaptation (LoRA) to ensure computational efficiency while maintaining high performance on the Kvasir-VQA-x1 dataset. Our approach utilizes a training split of 3,000 images and a validation split of 500 images, randomly sampled from the official training and test splits, respectively, using a fixed seed (42) for reproducibility. The subset selection, performed via `shuf`, balances computational constraints with dataset diversity, achieving a compact 630MB quantized model suitable for resource-constrained clinical environments.

3.1. Task overview and dataset

We addressed Subtask 1 of the Medico 2025 challenge [1], focusing on VQA for GI imaging. This subtask requires models to answer clinically relevant questions paired with GI images, testing multimodal reasoning capabilities. The Kvasir-VQA-x1 dataset comprises 159,549 QA pairs built on 6,500 GI images, extending Kvasir-VQA [5] and HyperKvasir [6]. It includes complexity scores to assess question difficulty, covering anatomical, diagnostic, and procedural queries. These scores enable targeted training on varied question types.

3.2. VQA pipeline

The PaliGemma-3B model integrates visual features from GI endoscopic images with tokenized clinical questions via cross-attention, generating answers for diverse question types (yes/no, single/multiple-choice, color, location, numerical). We applied LoRA (rank=128, alpha=64, dropout=0.05) to update all linear attention layers efficiently, reducing trainable parameters. Using the BitsAndBytes library (nf4 quantization, float16 compute dtype, double quantization enabled), we achieved a 630MB model, suitable for deployment on consumer-grade hardware. Answers align with challenge metrics: BLEU, ROUGE (1/2/L), METEOR and CHRF++.

3.3. Training configuration

We fine-tuned a 4-bit-quantized PaliGemma-3B model with LoRA adapters on 3,000 training images and validated on a 500-sample test split from the Kvasir-VQA-x1 dataset. Images were preprocessed to 224×224 resolution with normalization (mean=[0.5, 0.5, 0.5], std=[0.5, 0.5, 0.5]) and augmented with random rotations (±15 degrees), color jitter (brightness=0.2, contrast=0.2, saturation=0.2, hue=0.1), resized crops (scale 0.8–1.0), and horizontal flips. Using the `ms-swift` framework, training spanned 3 epochs (561 steps) with a peak learning rate of 1e-4 via a cosine scheduler (0.03 warmup ratio), per-device batch size of 2, gradient accumulation (8 steps, effective batch size 16), weight decay 0.05, and gradient checkpointing enabled. The best model, selected by minimum validation loss (0.4694 at step 500), was saved as a 630MB LoRA adapter in `output_Kvasir-VQA-x1/v0-20250913-181923/checkpoint-500`. Training ran ∼3 hours on two NVIDIA T4 GPUs (2×16 GB)..

3.4. Reproducibility details

We ensured reproducibility with a fixed seed (42) for dataset sampling and model initialization. The environment used Python 3.11, PyTorch 2.1+, `ms-swift`, `bitsandbytes` (0.41+), and `transformers` (4.35+), installed via `pip install ms-swift bitsandbytes wandb evaluate rouge_score`. Evaluation followed official MediaEval scripts for BLEU, ROUGE, METEOR and CHRF++. The training command was:

```
swift sft --dataset Kvasir-VQA-x1-train-3000.jsonl \
          --val_dataset Kvasir-VQA-x1-test-500.jsonl \
          --model google/paligemma-3b-pt-224 \
          --train_type lora \
          --quant_bits 4 \
          --lora_rank 128 \
          --lora_alpha 64 \
          --lora_dropout 0.05 \
          --num_train_epochs 3 \
          --learning_rate 1e-4
```

4. Results and Evaluation

We evaluate our LoRA-enhanced PaliGemma-3B model on Subtask 1 of the MediaEval Medico 2025 challenge for VQA on the Kvasir-VQA-x1 dataset. Performance is assessed on a 500-sample validation split from Kvasir-VQA-x1 full dataset and a private test set, using metrics: BLEU, ROUGE (1/2/L), METEOR and CHRF++.

Table 1 presents performance across complexity levels (1–3) and overall. On the validation set, BLEU is 0.3757, ROUGE-1 0.6357, and METEOR 0.6040. Complexity 3 leads (BLEU=0.4166), while Complexity 2 lags (BLEU=0.2969). The private test set shows BLEU=0.3973 and ROUGE-1=0.6465. Figure 1 illustrates the distribution of predicted visual feature attributes across complexity levels on the validation set. Each polygon represents the average activation of key visual categories—such as abnormalities, instruments, landmarks, and polyps—revealing that higher complexity questions (blue) involve richer combinations of visual cues compared to simpler ones (red and green).

Table 1
Performance metrics on the validation (Full) and private test sets across complexity levels for Subtask 1.

Set	Level	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CHRF++	BLEU
Full	Complexity 1	0.6512	0.4685	0.6423	0.6049	61.6168	0.4018
	Complexity 2	0.5911	0.3731	0.5630	0.5522	55.5914	0.2969
	Complexity 3	0.6639	0.4547	0.6079	0.6552	61.4107	0.4166
	Overall	0.6357	0.4328	0.6049	0.6040	59.4443	0.3757
Private	Complexity 1	0.6608	0.4586	0.6514	0.5963	60.8710	0.4165
	Complexity 2	0.6055	0.3865	0.5673	0.5714	55.6666	0.3121
	Complexity 3	0.6741	0.4869	0.6206	0.6725	62.3294	0.4421
	Overall	0.6465	0.4436	0.6136	0.6119	59.8024	0.3973

Ablation study: We conducted ablation studies to assess the impact of LoRA rank, quantization, and data augmentation on our LoRA-enhanced PaliGemma-3B model for the MediaEval

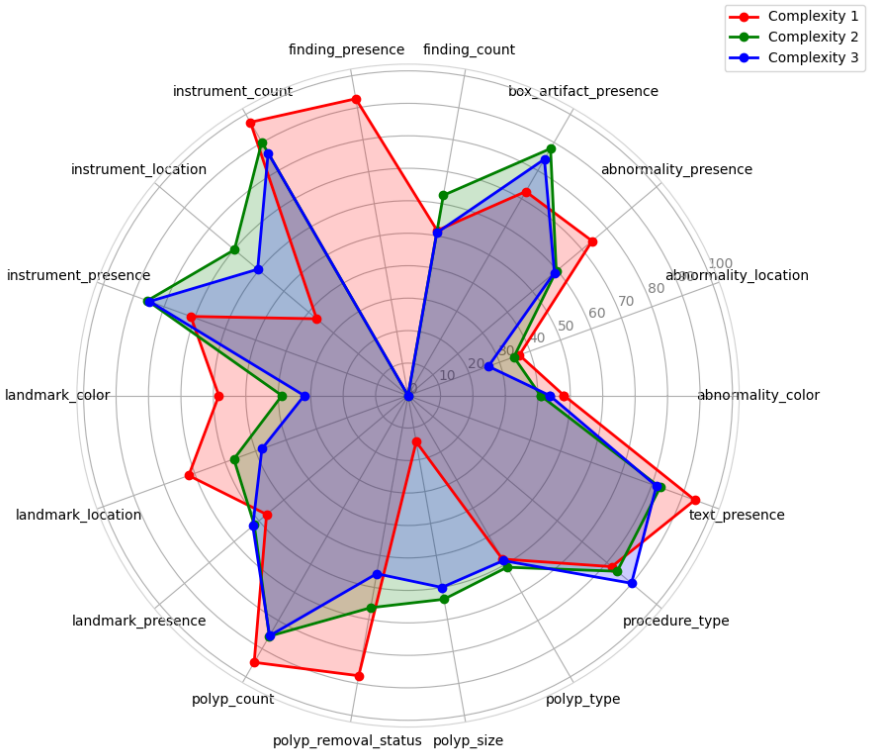

Figure 1: Radar plot illustrating the distribution of predicted visual feature attributes across different question complexity levels (1–3) on the validation set. Each axis represents a distinct visual attribute category such as *abnormality presence*, *instrument count*, *landmark location*, and *polyp characteristics*. The filled regions correspond to the average attribute frequencies per complexity level, highlighting the shift in visual reasoning requirements from simpler (red) to more complex (blue) questions.

Medico 2025 VQA challenge. Testing LoRA rank = 128 achieved BLEU score = 0.3757 and ROUGE-L = 0.6049. Removing 4-bit quantization increased model size to 2.8GB with minimal metric gains (+0.01 BLEU). Excluding augmentation (rotations, jitter, flips) decreased BLEU by 8%, confirming its role in robust VQA.

5. Discussion and outlook

PaliGemma-3B, our LoRA-enhanced VQA model in GI imaging has high performance on the private test set of the Kvasir-VQA-x1 dataset (BLEU: 0.3973, ROUGE-L: 0.6136, METEOR: 0.6119), especially on Complexity 3 questions (BLEU: 0.4421, ROUGE-L: 0.620). The 630MB model, which is made possible by LoRA and quantization is ideal to be used in clinical settings but it might be problematic because of its potential to overfit to the 3,000-sample training split and the 500-sample validation set, which is small enough to restrict the generalization of the model to various clinical situations. Also, the model is resource-efficient, but its implementation requires high compute devices. Our future work will be focusing on enhancing reasoning for Complexity 2 questions, expanding the training dataset beyond 3,000 samples, and validating on larger, more diverse sets to improve robustness and clinical applicability.

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, 2025.

[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, 2025, pp. 53–63.

[3] J. J. Lau, S. Gayen, A. Ben Abacha, et al., A dataset of clinically generated visual questions and answers about radiology images, ArXiv e-prints (2018).

[4] X. He, Y. Zhang, L. Mou, et al., PathVQA: 32,000 open-ended questions for pathology visual question answering, ArXiv e-prints (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, 2024, pp. 3–12.

[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, et al., HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, Sci. Data 7 (2020) 1–14.

[7] 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, 2024, pp. 140–164.

[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, 2025, pp. 398–406.

[9] E. J. Hu, Y. Shen, P. Wallis, et al., LoRA: Low-Rank Adaptation of Large Language Models, ArXiv e-prints (2021).

[10] M. Chaichuk, S. Gautam, S. Hicks, E. Tutubalina, Prompt to Polyp: Medical Text-Conditioned Image Synthesis with Diffusion Models, ArXiv e-prints (2025).