Curriculum-Guided Fine-Tuning for Multimodal VQA in GI Endoscopy (Team Lama4Vision)

Mahdi Azmoodeh-Kalati[1], Mohammad Sadegh Maghareh[2], Saba Alavi[3] and Reza Lashgari[1,*]

[1]*Institute of Medical Science and Technology, Shahid Beheshti University, Tehran, Iran*
[2]*Computer Science Department, Amirkabir University of Technology, Tehran, Iran*
[3]*Artificial Intelligence Department, Shahid Beheshti University, Tehran, Iran*

Abstract

Visual Question Answering (VQA) for gastrointestinal (GI) endoscopy images presents unique challenges due to the complex visual features and domain-specific language involved in medical questions and answers. In this work, we investigate a multimodal fine-tuning strategy to adapt a modern vision–language architecture to the Kvasir-VQA-x1 dataset. Our approach incorporates curriculum-guided fine-tuning, where the model is progressively trained on easy, medium, and hard question–answer pairs, enabling smoother convergence and improved reasoning over diverse question types. To enhance robustness under acquisition variability, we employ clinically safe weak data augmentation that preserves diagnostic semantics while improving generalization. We apply lightweight, parameter-efficient tuning via LoRA adapters and adopt a staged training schedule to balance resource constraints while maintaining strong generalization capabilities. Experimental results on an internal validation subset show consistent improvements across BLEU, ROUGE, and METEOR, confirming the benefits of curriculum-guided and augmentation-aware fine-tuning for multimodal medical reasoning.

1. Introduction and RelatedWork

Gastrointestinal (GI) diseases such as colorectal and gastric cancers pose a major global health burden, where early detection via endoscopy is vital for improving patient outcomes. Endoscopic images are often complex and affected by artifacts (e.g., glare, blur), making interpretation challenging and operator-dependent. Computer-assisted methods have thus been explored to enhance diagnostic accuracy and reduce observer variability [1, 2, 3]. While our earlier work investigated ensemble-based CNN architectures for histopathological image classification [4], the present study broadens this line of research toward multimodal vision–language reasoning in gastrointestinal endoscopy. Early research in GI endoscopy used convolutional neural networks (CNNs) for lesion detection and classification, achieving promising yet limited generalization [5, 6]. Later architectures such as EfficientNet and ensemble methods improved diagnostic performance on Kvasir datasets [7, 8]. Beyond detection, Medical Visual Question Answering (MedVQA) extends image analysis to natural-language reasoning [9], enabling models to answer clinical questions (e.g., "How many polyps are visible?"). Datasets such as Kvasir-VQA [10] and Kvasir-VQA-x1 [11] benchmark multimodal vision–language models, covering diverse question types and complexity levels. Parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA) enable efficient adaptation of large vision–language models (VLMs) under limited resources, yet remain underexplored in MedVQA. This work proposes a multimodal fine-tuning strategy integrating LoRA-based adaptation, curriculum-guided

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

[*] Corresponding author.

r_lashgari@sbu.ac.ir (R. Lashgari)

© 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)

learning, and clinically safe weak augmentation. Our contributions for the MediaEval 2025 MedVQA task [12] are:

1. Evaluating LLaMA-3.2-Vision and Qwen2-VL 2B on Kvasir-VQA-x1;
2. Analyzing parameter-efficient fine-tuning via LoRA; and
3. Employing curriculum-guided fine-tuning to improve robustness and generalization across question complexities.

2. Methodology

2.1. Dataset Split and Leakage Analysis

We used the Kvasir-VQA-x1 dataset, comprising 159,549 question-answer pairs grounded on approximately 6,449 unique gastrointestinal endoscopy images. Datails of split has been mentioned in Table 1 The original dataset split was defined at the QA-pair level, causing substantial leakage issues: 3,821 images (about 59.2% of unique images) appeared in both training and test subsets, violating the assumption of unseen test data.[10][11]

2.2. Base Model and Zero-Shot Baseline

We evaluated recent vision–language models (e.g., LLaMA–3.2–Vision, Qwen2–VL, Pixtral) under limited compute. **Qwen2–VL 2B–Instruct (bnb 4–bit)** was selected as the primary base model due to its memory efficiency and competitive performance, while LLaMA–3.2–Vision 11B served only for preliminary comparison. Before fine-tuning, a **zero-shot** evaluation was conducted on an internal subset of Kvasir-VQA-x1, where each image–question pair was prompted without task-specific training. As expected, zero-shot performance was moderate—adequate for simple identification but weak on clinically complex reasoning—highlighting the need for domain-specific adaptation.

2.3. Parameter-Efficient Fine-Tuning (LoRA)

Full fine-tuning of large vision–language models is computationally expensive and prone to overfitting. We therefore employ **Low-Rank Adaptation (LoRA)**, a parameter-efficient method that freezes the backbone and learns rank-r adapters ($\Delta W = BA$, where $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$, scaled by α/r). We adopt a compact configuration ($r = \alpha = 16$), following best practices for QLoRA on 4-bit quantized models, achieving a balance between efficiency and stability. A progressive learning rate schedule ($5 \times 10^{-5} \rightarrow 3 \times 10^{-5} \rightarrow 2 \times 10^{-5}$) aligns with the curriculum setup—higher at Stage 1 for "Easy" data and gradually reduced for "Moderate" and "Hard" examples to prevent catastrophic forgetting and improve generalization. Training uses a batch size of 512 (128 per device × 4 accumulation steps) with Unsloth's 8-bit AdamW optimizer, providing an effective trade-off between convergence and GPU constraints.

2.4. Incremental Training by Complexity

We employ a curriculum learning strategy on the Kvasir–VQA–x1 dataset, organizing samples by question complexity. As summarized in Table 1, fine-tuning proceeds in three one-epoch stages using LoRA (with the backbone frozen), where weights are carried forward and checkpoints saved after each stage.

Stage 1 (L1–Easy). Train on level–1 Q&A pairs to learn basic visual cues and direct mappings.

Table 1
Fine-tuning configuration and explicit curriculum details on Kvasir-VQA-x1. The table summarizes all hyperparameters, LoRA setup, augmentation strategies, and the three-stage curriculum schedule (Easy, Moderate, Hard) described in Section 2.4. These settings ensure reproducibility and transparency of the training pipeline.

Hyperparameter	**Value**
Base model	Qwen2–VL 2B–Instruct (bnb 4–bit)
LoRA rank (r)	16
LoRA α	**16**
LoRA Dropout	0.0
LoRA Bias	none
Random State	3407
LoRA Target Modules (Unsloth Configuration)	
Finetune Vision Layers	**True**
Finetune Language Layers	**True**
Finetune Attention Modules	**True**
Finetune MLP Modules	**True**
Weak Data Augmentation Parameters (Applied in All Stages)	
Image Size	224*224
Global Random Seed	42
RandomResizedCrop Scale	(0.9, 1.0)
RandomResizedCrop Ratio	(0.95 × AR, 1.05 × AR)
RandomRotation Degrees	(-10, 10)
RandomAffine Translate	(0.1, 0.1)
ColorJitter Brightness/Contrast	0.2 / 0.2
Curriculum Schedule and Training Details	
System	single A100 GPU (VRAM 40GB), 128 core CPU , 128 GB RAM
Curriculum stages	3 (Easy; Easy+Medium; All)
Epochs per stage	1
Total Training Samples	143,594
Stage 1 Samples (Easy)	49,360
Stage 2 Samples (Easy+Moderate)	96,458
Stage 3 Samples (All)	143,594
Per-device batch size	128
Gradient accumulation	4
Total Effective Batch Size	**512**
Learning rate (Stage 1/2/3)	5×10^{-5} / 3×10^{-5} / 2×10^{-5}
LR Scheduler Type	**linear**
Warmup steps	**2**
Logging steps	**5**
Optimizer	AdamW (8-bit)
Experiment Tracking	**wandb**

Stage 2 (L1+L2–Easy+Moderate). Continue from Stage 1 using levels 1–2; mixing L1 data mitigates forgetting and introduces moderate reasoning.

Stage 3 (L3–Hard). Continue from Stage 2 using level–3 samples to enhance complex, clinically grounded reasoning. LoRA's adapter-only updates help retain earlier knowledge while specializing to difficult cases.

3. Results and Evaluation

We evaluate the model on the Kvasir–VQA–x1 test split using standard natural language generation metrics for question answering. The evaluation focuses on both linguistic fidelity and semantic consistency between generated and reference answers. *BLEU* [13] measures n-gram precision (we report BLEU-4 with brevity penalty); *METEOR* accounts for synonym and stem-level matches, giving partial credit for semantically related words; and *ROUGE* (ROUGE-1 and ROUGE-L F-measure) computes overlap in unigrams and the longest common subsequence. Together, these metrics capture both surface-level accuracy (*BLEU*, *ROUGE*) and semantic adequacy (*METEOR*), offering a balanced evaluation of the model's language generation performance.

3.1. Quantitative Results

LoRA fine-tuning improved overall model performance on Kvasir–VQA–x1 (Task 1) To provide clearer quantitative insight, Table 2 shows scores per question complexity level.

Table 2
Performance by question complexity (public split).

Complexity	BLEU-4	METEOR	ROUGE-1	ROUGE-L
Level 1	0.2739	0.5484	0.4806	0.4704
Level 2	0.2277	0.5877	0.5873	0.5546
Level 3	0.3327	0.5785	0.6292	0.5696

Compared to the best system, which achieved scores of BLEU 0.4975, ROUGE-1 0.7261, ROUGE-2 0.5476, ROUGE-L 0.7, and METEOR 0.6988, our model's results (BLEU 0.2739, ROUGE-1 0.5548, ROUGE-2 0.3367, ROUGE-L 0.5242, METEOR 0.5093) show a gap. This gap mainly stems from current resource constraints and limitations of the chosen model architecture, which restricted extensive experimentation and optimization. Nonetheless, our approach benefiting from domain adaptation and curriculum learning is promising.

3.2. Qualitative Insights

Beyond quantitative scores, qualitative inspection of model outputs provides deeper understanding of system behavior. The fine-tuned models consistently produce accurate answers for visual identification tasks (e.g., detecting polyps, distinguishing anatomical regions), and curriculum learning notably improves consistency on complex questions. However, descriptive or reasoning-intensive questions (e.g., "What treatment is needed?") remain challenging, occasionally leading to *hallucinations*—the generation of medical terms not present in the image. Failure cases often arise in images with strong artifacts (e.g., glare, blur) or multiple abnormalities, where responses may be partially correct (e.g., accurate detection but incomplete reasoning).

Overall, LoRA fine-tuning substantially boosts performance over zero-shot baselines, while the curriculum-guided approach enhances robustness across complexity levels and helps retain performance on easier questions without catastrophic forgetting on harder ones.

4. Limitations and Future Work

The final configuration is considered optimal under resource constraints rather than a global optimum, due to limited compute (single A100 GPU). Hyperparameters were adopted from the Unsloth library as a practical baseline.

Future Work — Leakage Mitigation. We identified significant image-level overlap (≈59%) between training and test splits in Kvasir–VQA–x1, indicating visual leakage. Future work will adopt a strict image-level split, assigning each image and its Q&A pairs exclusively to one split, to ensure fair evaluation and eliminate leakage.

Declaration on Generative AI

ChatGPT was used to improve grammar and structure. The authors reviewed and edited all content and take full responsibility for the final manuscript.

Implementation scripts and Model

Resources: GitHub Repository and Kaggle Notebook, Hugging Face Model, Kaggle Dataset.

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) 283.

[2] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. De Lange, P. T. Schmidt, H. D. Johansen, et al., Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, in: International Conference on Multimedia Modeling, Springer, 2021, pp. 218–229.

[3] G. Parasher, M. Wong, M. Rawat, Evolving role of artificial intelligence in gastrointestinal endoscopy, World journal of gastroenterology 26 (2020) 7287.

[4] M. Azmoodeh-Kalati, H. Shabani, M. S. Maghareh, Z. Barzegar, R. Lashgari, Leveraging an ensemble of efficientnetv1 and efficientnetv2 models for classification and interpretation of breast cancer histopathology images, Scientific Reports 15 (2025) 21541.

[5] M. J. Kim, S. H. Kim, S. M. Kim, J. H. Nam, Y. B. Hwang, Y. J. Lim, The advent of domain adaptation into artificial intelligence for gastrointestinal endoscopy and medical imaging, Diagnostics 13 (2023) 3023.

[6] C. Zhang, L. Wu, The application of artificial intelligence in gastrointestinal endoscopy: a state-of-the-art review, Journal of Digital Health 1 (2022) 3–18. URL: https://ojs.luminescience.cn/JDH/article/view/42. doi:10.55976/jdh.120221423-18.

[7] R. M. Patil, S. Giripunje, Deep learning-based detection and classification of gastrointestinal tract diseases in endoscopy images, in: 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI), IEEE, 2024, pp. 1–6.

[8] S. Siddiqui, J. A. Khan, S. Algamdi, Deep ensemble learning for gastrointestinal diagnosis using endoscopic image classification, PeerJ Computer Science 11 (2025) e2809.

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

[10] S. Gautam, A. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, M. A. Riegler, Kvasir-vqa: A text-image pair gi tract dataset, in: Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications (VLM4Bio '24), ACM, 2024, p. 10 pages. doi:10.1145/3689096.3689458.

[11] 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.

[12] 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.

[13] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of ACL, 2002, pp. 311–318.