Enhancing encoder-decoder architecture to visual question answering task for gastrointestinal images

Minh-Triet Bui[1,2,3], Nam-Tran Nguyen-Thuy[1,2,3], Hai-Dang Nguyen[1,2,*] and Minh-Triet Tran[1,2,*]

[1] *University of Science, VNU-HCM*

[2] *Vietnam National University, Ho Chi Minh City, Vietnam*

[3] *VNU-HCM High School for the Gifted*

Abstract

Med-VQA is an expanding field, but most research so far has centered on few fields such as radiology, with comparatively little focus on gastrointestinal imaging (GI). The task of answering visual questions for gastrointestinal images aims to utilize LLMs in a specified field medical VQA task. In addressing this challenge, we propose our encoder-decoder architecture approach. Additionally, we investigated whether the encoder part actually helps extract relevant features or not. The architecture achieved an accuracy of 73.18%, demonstrating performance comparable to recent causal language models despite performing worse on long, complex questions where long dependencies emerge. Further optimization with modern techniques could enhance its effectiveness.

1. Introduction and Related Work

Gastrointestinal diseases, including colorectal cancer, represent a significant global health burden and demand early, accurate, and reliable diagnostic approaches. Despite the growing number of medical visual question answering (MedVQA) models developed to assist in clinical interpretation, their adoption in gastrointestinal imaging has not been widespread compared to other fields like radiology or cardiology. To address this challenge, we propose a novel VQA framework for the MediaEval 2025 Medico challenge [1]. Our model utilizes **InstructBLIP (FlanT5$_{\text{XXL}}$)** which was selected through an extensive experimentation process.

The most popular pre-trained architecture for image-text-to-text tasks is decoder-only since it is well-adapted to many tasks with its flexibility. Early efforts for medical-oriented visual question answering tasks for LLM such as LLavaMed [2] from Microsoft and MedGemma [3] from Google have proven their potential. The presence of encoder-decoders in these types of tasks is relatively less but has been researched, for example this paper about image analysis[4]. Our experimental process with different models was heavily influenced by the original paper's model choices for testing[5].

MediaEval'25: Multimedia Evaluation Workshop

[*] Corresponding author.

[†] These authors contributed equally.

student240238@ptnk.edu.vn (M. Bui); student240237@ptnk.edu.vn (N. Nguyen-Thuy); nhdang@selab.hcmus.edu.vn (H. Nguyen); tmtriet@fit.hcmus.edu.vn (M. Tran)
0000-0003-0888-8908 (H. Nguyen); 0000-0003-3046-3041 (M. Tran)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Methodology

2.1. Selecting model

We decide to use an encoder-decoder, specifically T5 [6] which is rarely experimented with. We investigated BLIP2 (Bootstrapping Language-Image Pre-training 2) [7] where a *"bridge"* called **Q-Former** smooths the connection of a vision encoder to a LLM. InstructBLIP [8] is an updated version and shows improvements. Eventually, we decided to fine-tune an encoder-decoder which is **InstructBLIP (FlanT5$_{\text{XXL}}$)**. Our implementation is publicly available at [9].

2.2. Training strategy

We randomly selected 10% of the training set for evaluation during the training phase. The test set is reserved for final evaluation. For an efficient fine-tuning process, we decided to use LoRA [10] because it reduces the number of trainable parameters. Hyperparameters are shown in Table 1. We applied a linear decay learning rate scheduler per step with the starting learning rate at 5e-5. An additional Lora dropout was applied to prevent overfitting. We chose a large α as the task is quite different from pre-trained tasks.

Table 1
Technical specifications

Params	Dtype	Epochs	Effective batch size	Lr	Lr scheduler	Lora r/α	Lora dropout
12.3B	bfloat16	10	64	5e-5	Linear decay	16 / 48	0.05

We consider medical images to be sensitive to semantic distortion, so no augmentation was applied. Furthermore, we did not give instructions to the model, since both models with and without instructions had similar results. We set our maximum token length for questions and answers at 55. A linear decay learning rate scheduler was applied for smoother convergence. We tested and concluded that 10 epochs worked just fine without the risk of overfitting. The loss is not reported as it may differ depending on the framework.

We seed everything to 27022009. The train set and the validation set were split using the train_test_split function from the sklearn library with a fixed seed. The model was fine-tuned on an A100-40GB GPU.

3. Results

3.1. NLP Metrics

Table 2
NLP Metrics compare to [5]

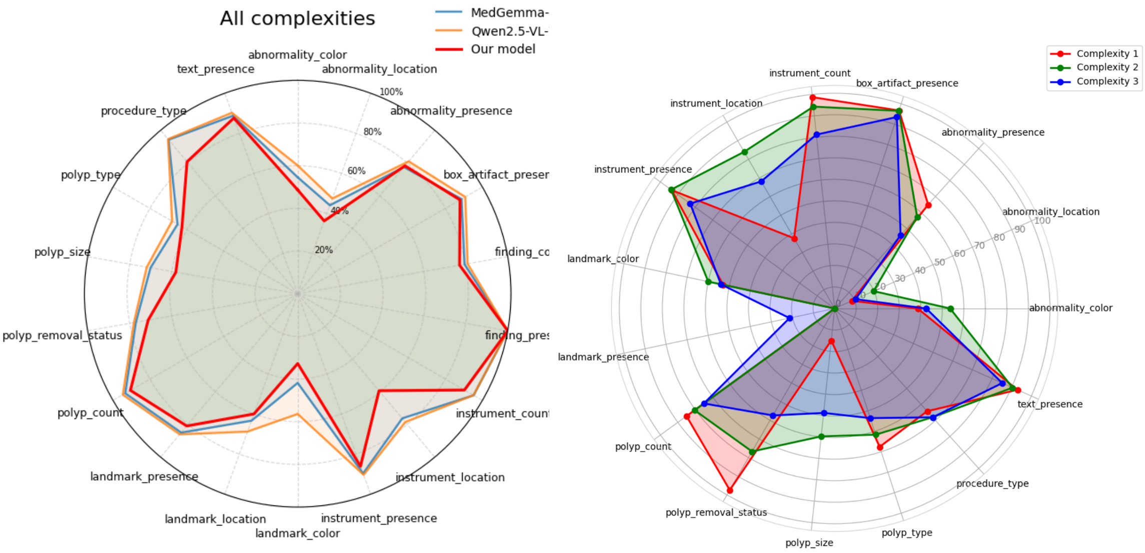
Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CHRF++	BLEU	BERTScore F1
MedGemma-4B	0.677	0.476	0.648	0.648	63.93	0.428	0.948
Qwen2.5-VL-7B	**0.716**	**0.534**	**0.690**	**0.689**	**67.91**	**0.478**	**0.954**
InstructBLIP (FlanT5$_{\text{XXL}}$)	0.706	0.530	0.683	0.656	63.13	0.432	0.948

From table 2, Qwen2.5-VL with a more modern design (2025) and a training strategy surpasses the rest. On the other hand, our model, with more parameters, is competitive with MedGemma, which is a medical fine-tuned version of gemma-4B and also a newer model (2025). These models were trained on hyperparameters reported in Table 1.

3.2. Accuracy

Model's accuracy is judged based on Qwen3-30B-A3B [11].

Figure 1: Public vs Private



(a) Public results compare with other models

(b) Private results

In figure 1a, it can be shown that InstructBLIP nearly equalizes others in a few categories despite performing worse on the remaining features. This may be caused by the fact that compared to decoder-only models, encoder-decoders have a weaker decoder to generate an answer based on long dependencies. Overall, all models have the same accuracy pattern.

On the other hand, figure 1b shows an interesting pattern in which some categories achieve a high accuracy while the others are dramatically lower, especially the 2 features landmark_presence and abnormality_location. This may be due to different lighting conditions or other environmental factors.

Table 3
Accuracy

Model	Overall	Complexity 1	Complexity 2	Complexity 3
InstructBLIP	73.18%	76.16%	78.00%	68.89%
Qwen	**80.62%**	**80.95%**	**81.79%**	**79.74%**
MedGemma	78.62%	77.46%	80.39%	77.86%

Table 3 illustrates accuracy per complexity on public test, it shows Qwen dominating in performance. Nevertheless, InstructBLIP still performs well, highlighting the potential of encoder-decoder models in general.

3.3. Private test result

Table 4
Private test results

Level	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CHRF++	BLEU	BERTScore (P/R/F1)
Complexity 1	**0.67**	**0.50**	**0.67**	0.61	**63.08**	0.35	**0.95 / 0.94 / 0.94**
Complexity 2	0.64	0.44	0.61	**0.62**	59.62	0.35	0.94 / 0.94 / 0.94
Complexity 3	0.67	0.50	0.62	0.61	57.71	**0.41**	0.94 / 0.93 / 0.94
Overall	0.66	0.48	0.63	0.61	59.32	0.40	0.94 / 0.94 / 0.94

Table 4 indicates slightly worse performance compared to the public test results. This may be due to a more challenging private test set. However, the results are still acceptable.

4. Discussion and Outlook

In conclusion, our research illustrates the potential of an encoder-decoder architecture, which is relatively formidable if it is further optimized and applied to current state-of-the-art methods and techniques. Encoder-decoders are ideal for many non-interactive and specialized visual question answering tasks as encoders help to extract information. Additionally, in short questions, which have only one sub-question, the model tends to perform very well, but it performs worse in questions with multiple sub-questions. Further solutions may emphasize more on the decoder part for its capacity to generate an answer on long dependencies as they are in long questions.

5. Generative AI Use Declaration

This paper was assisted with the use of Gen AI, specifically ChatGPT and Overleaf's autocomplete feature, to enhance vocabulary and grammar for improved domain specificity and overall a better reading experience.

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, Dublin, Ireland and Online, 2025. doi:10.48550/arXiv.2508.10869.

[2] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: Training a large language-and-vision assistant for biomedicine in one day, arXiv preprint arXiv:2306.00890 (2023).

[3] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, et al., Medgemma technical report, arXiv preprint arXiv:2507.05201 (2025).

[4] H. Aboutalebi, M. Samadzadegan, W. C. C. Kwan, Medusa: Multi-scale encoder-decoder self-attention deep neural network architecture for medical image analysis, arXiv preprint arXiv:2110.06063 (2021). arXiv:2110.06063.

[5] 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.

[6] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. F. et al., Scaling instruction-finetuned language models, 2022. URL: https://arxiv.org/abs/2210.11416.

[7] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL: https://arxiv.org/abs/2301.12597. arXiv:2301.12597.

[8] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL: https://arxiv.org/abs/2305.06500. arXiv:2305.06500.

[9] lucky-and minie222702, Mediaeval2025—medico, https://github.com/lucky-and-minie222702/MediaEval2025---Medico, 2025. Accessed: 2025-10-20.

[10] 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, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[11] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Z. et al., Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388.