Towards Better Gastrointestinal Diagnosis: Evaluating Vision-Language Models For GI VQA

Notebook for the ImageCLEFmedical Lab at CLEF 2025

Omar Adjali[1]

[1]*Paris-Saclay University, Gif-sur-Yvette, France*

Abstract

Gastrointestinal (GI) image analysis is critical for early diagnosis and treatment of GI diseases, which remain a leading cause of global morbidity and mortality. Visual Question Answering (VQA) in medical imaging enables interpretable and interactive AI systems to support clinical decision-making. This paper presents our submission to the ImageCLEFmed 2025 MedVQA task, which targets medical VQA on gastrointestinal endoscopic images using the Kvasir-VQA dataset. We evaluate two primary approaches: (1) a multimodal Chain-of-Thought (CoT) reasoning framework that decomposes questions into interpretable reasoning steps, and (2) a simple fine-tuning strategy on large-scale generative models. Extensive experiments across multiple vision-language models, including Qwen2-VL and BLIP2, show that fine-tuning significantly outperforms CoT in both validation and test settings. Our best-performing model, achieves a METEOR score of 50 on the test set. We also carried out qualitative and quantitative analysis to demonstrate the strengths and weaknesses of our best performing approach, and hence suggesting some insights to tackle the most challenging aspects in the Kvasir-vqa task.

Keywords

Medical VQA, ImageCLEFmed 2025, Multimodal AI, Clinical Question Answering, Synthetic GI Images

1. Introduction

Gastrointestinal (GI) image analysis is crucial due to the high prevalence and mortality of GI diseases, particularly cancers, which account for millions of new cases and deaths globally each year. Early and accurate detection through endoscopic imaging can significantly reduce these numbers, yet human performance variability often leads to diagnostic oversights, such as the 20% polyp miss-rate in colonoscopies [1]. With the unprecedented development in Artificial Intelligence (AI) systems, AI-enabled medical support systems offer a promising solution by enhancing diagnostic accuracy and consistency. In particular, Visual Question Answering (VQA) serves as a vital interface, allowing experts to interact with medical images through natural-language queries, thereby improving interpretability and decision-making. Furthermore, synthetic data generation plays a pivotal role in overcoming the data scarcity challenge in medical AI by augmenting datasets with realistic and diverse examples that support robust model training. Similar to previous initiatives such in [2], the ImageCLEFmed 2025 MedVQA challenge [3] contributes to address the aforementioned challenges using the recently introduced HyperKvasir dataset [1], which serves as a benchmark for addressing the task of Visual Question Answering on gastrointestinal data and GI synthetic image generation.

In this paper, we report our findings in addressing the Medical Visual Question Answering (MedVQA) task. It aims to generate accurate answers to clinical queries based on medical images. In particular, the ImageCLEFmed 2025 MedVQA task is of high relevance for improving gastrointestinal diagnostic processes.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

adjali.omar@gmail.com (O. Adjali)

0000-0002-6021-7776 (O. Adjali)

2. Related Work

2.1. MedVQA Approaches

MedVQA has gained significant attention as a critical task in biomedical AI, requiring models to generate accurate textual answers conditioned on visual medical images. Early MedVQA Approaches addressed tasks with limited annotated data. For example, [4] proposed a framework combining Convolutional Denoising Auto-Encoders (CDAE) and Model-Agnostic Meta-Learning (MAML) to utilize both unlabeled data and few-shot learning. [5] further introduced a conditional reasoning approach that adapts reasoning strategies based on question types (e.g., closed- vs. open-ended) which significantly improved performance on the VQA-RAD dataset [6]. To manage the diversity of question types, [7] also proposed CGMVQA, a hybrid model that handles both classification and generative answering via transformer-based architecture. Further works have employed contrastive learning for better visual feature extraction in low-data regimes. In particular, [8] proposed a dual approach combining a reasoning module with a contrastively trained visual encoder. Similarly, [9] fine-tuned CLIP on PubMed image-text pairs, showing notable improvements over visual-only pretrained models through the introduction of the PubMedCLIP encoder. More recently, the generative paradigm has gained interest with the emergence of Large Vision-Language Models (LVLM). [10] introduced PMC-VQA, a large-scale dataset comprising over 227k image-question-answer pairs, and proposed a generative model fine-tuned for free-form answering. Similarly, [11] presented LLaVA-Med, trained using a novel curriculum learning strategy on instruction-following data generated by GPT-4, outperforming previous supervised approaches in both accuracy and versatility. In order to improve interpretability which is crucial for clinical applications, recent work leveraged self-reflexion reasoning enabled by large language models (LLMs). For example, [12] proposed MedCoT, which relies on a multi-expert diagnostic collaboration through hierarchical Chain of thought and Mixture of Experts. [13] introduced MedThink, which integrates Medical Decision-Making Rationales (MDMRs) into a generative model to make the reasoning process transparent and clinically verifiable.

2.2. MedVQA Datasets

The development of robust and clinically relevant Visual Question Answering (VQA) systems for medicine is heavily dependent on high-quality annotated datasets. Over the past few years, several notable datasets have emerged, each addressing unique aspects of medical image understanding through natural language queries. VQA-RAD [14] is the first manually curated medical VQA dataset tailored to radiology. It comprises over 3,500 natural question-answer (QA) pairs covering 315 unique radiological images. The questions were authored by clinical trainees with medical imaging experience, ensuring medical realism. Similarly, [15] introduced PathVQA, the first VQA dataset focused on pathology including open-ended and yes/no questions. More recently, [16] proposed SLAKE, a large bilingual dataset that covers more body parts with rich semantic labels annotated by experienced physicians. In the context of ImageCLEFmed 2025 MedVQA challenge, [17] proposed the Kvasir-VQA dataset which extends the HyperKvasir and Kvasir-Instrument datasets by introducing over 52,000 synthetic question-answer pairs for 6,500 images across various gastrointestinal findings, including polyps, esophagitis, and ulcerative colitis. These QA pairs encompass a range of formats such as yes/no, multiple choice, location, and numerical count, and were validated by medical experts. This dataset targets image captioning, diagnostic VQA, and synthetic image generation, enabling research in GI tract diagnostics and fine-grained instrument recognition. It also supports training generative models such in [18] for image synthesis based on medical prompts. Finaly, most recentl, [19] proposed OmniMedVQA, a new large-Scale comprehensive benchmark for evaluating large vision-language models in the medical domain. It comprises 118,010 real medical images and 127,995 question-answer (QA) pairs, collected from 73 distinct datasets, spanning 12 imaging modalities (e.g., MRI, CT, X-Ray, Ultrasound) and over 20 human anatomical regions. OmniMedVQA QA pairs are systematically constructed to evaluate five major medical reasoning capabilities: modality recognition, anatomy identification, disease diagnosis, lesion grading, and biological attributes.

3. Task Overview and Dataset

3.1. Task Formulation

The Medical Visual Question Answering (MedVQA) task aims to develop models that can accurately answer clinically relevant questions about gastrointestinal (GI) endoscopic images. Leveraging the Kvasir-VQA dataset, the task combines computer vision and natural language understanding to simulate expert-level diagnostic reasoning. Formally, given an input medical image I and a natural language question Q associated with the image, the objective is to map the image-question pair to a natural language answer A that is accurate and contextually grounded in the image.

3.2. Kvasir-VQA Dataset

Table 1
Distribution of image categories and sources in the Kvasir-VQA dataset.

Image Category	Number of Images	Source Dataset
Normal	2500	HyperKvasir
Polyps	1000	HyperKvasir
Esophagitis	1000	HyperKvasir
Ulcerative Colitis	1000	HyperKvasir
Instrument	1000	Kvasir-Instrument
Total	**6500**	

The Kvasir-VQA dataset includes the following structured question types:

- **Yes/No:** Binary or ternary decisions (e.g., "Have all polyps been removed?").
- **Single-Choice:** Classification from a predefined list (e.g., "What type of polyp is present?" with Paris classification options).
- **Multiple-Choice:** Multi-label decisions (e.g., "Are there any abnormalities?" with options like polyp, esophagitis, etc.).
- **Color-Based:** Selection based on observed color (e.g., "What color is the abnormality?" with options like pink, red, etc.).
- **Location-Based:** Region selection (e.g., "Where in the image is the instrument?" with grid-based location options).
- **Numerical Count:** Estimation (e.g., "How many polyps are present?" with integer output).

4. Methodology

In this paper, we propose exploring two approaches to tackle the ImageCLEFmed 2025 MedVQA (task 1). We first investigate how a multimodal chain of thoughts (CoT) system would perform on the Kvasir-vqa dataset. Then, we evaluate a simple finetuning strategy using the kvasir-vqa training dataset and other medical training data.

4.1. Multimocal CoT

Chain-of-Thought (CoT) reasoning enables Large Language Models (LLMs) to explicitly decompose complex questions into intermediate reasoning steps [20, 21]. As shown in [12], MedVQA queries often require multi-step inference that combines clinical knowledge with image interpretation, thus, logical paths can be traced from questions to the final answer. Such a structured decomposition may help mitigate hallucinations and improve answer generation accuracy. Inspired from [12], we model the ImageCLEFmed 2025 MedVQA task using a multimodal CoT system. Given a Large Vision Languge Model (LVLM) and a question-image input pair (q,i), we perform the following inference steps: 1)

We generate a preliminary reasoning rationale *R* such that: *R* = LVLM(q, i, *P*), where *P* is the prompt instruction used to generate the rationale *R*. *P* is formulated as follows:

Rationale Instruction Prompt

You are a Vision Language Model assistant which helps an experienced doctor interpreting accurately interpreting and answering clinical questions based on gastrointestinal images. Given the image, provide a reasonable rationale for the question: {**QUESTION**}. Please proceed with a step-by-step analysis and provide a rationale.

Subsequently, *R* is used to generate the final answer *A*, such that: *A* = LVLM(q, i, *R*). We relied on the following prompt:

Answer Generation Prompt

You are a Vision Language Model assistant which helps an experienced doctor interpreting accurately interpreting and answering clinical questions based on gastrointestinal images. Given the image and the rationale: {**RATIONALE**}, Answer briefly the question: {**QUESTION**} with no extra text, rationales or explanation.

Since the generated rationale can be ineffective with regard to the ground truth answer A^*, we trained the LVLM on answering Kvasir-vqa questions in order to reduce discrepancies between rationales and the actual answers. The LVLM is trained on the following cross-entropy loss:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T} \log P(A^*, q, i, R) \tag{1}$$

where A^* is the ground truth answer, q is the question, i is the image, and R is the rationale.

4.2. Finetuning strategy

In the second approach, we performed answer generation using a generative model denoted G(·) with parameters Φ. Given a question q_j and the associated image i_j, the answer generator G is trained on the following cross-entropy loss over a batch of N question-image pairs:

$$\mathcal{L}_{\text{G}} = -\sum_{j=1}^{N} \log p_{\Phi}(s_j^* \mid q_j, i_j) \tag{2}$$

Where q_j is the j-th question, i_j is the image associated with q_i, s_j^* is the ground truth answer string for (q_j, i_j), $p_{\Phi}(s_j^* \mid q_j, i_j)$ is the probability of generating the correct answer from the text-image pair, Φ are the parameters of the multimodal answer generator.

4.3. Training details

We implemented all our experiments in Pytorch [22] and we relied on the Qwen2-VL-72B-instruct [23] LVLM for generating reasoning rationales. Afterward, we performed the CoT and finetuning training stages on LVLMs of different size and architectures: Qwen2-VL-7B-instruct, Qwen2-VL-32B-instruct [23] and BLIP2-Flan-T5-XL [24]. We trained for 10 epochs using a batch size of 4 and a learning rate of 2e-5 on a single A100 GPU. Throughout all finetuning experiments, the LVLMs are trained using LoRA [25] for efficient parameters optimization with the following configurations: $\{r = 8, lora\_alpha = 32, lora\_dropout = 0.1\}$ with BLIP2-Flan-T5-XL and $\{r = 8, lora\_alpha = 16, lora\_dropout = 0.05\}$ for Qwen's models. At inference time, decoding is performed using 3 beams search. Model checkpoint selection was done based on validation meteor performance.

5. Results and Discussion

We evaluated both Chain-of-Thought (CoT) and fine-tuned (FT) models using BLEU, ROUGE, and METEOR scores. The Qwen2-VL and BLIP2 model architectures were evaluated for both methodologies. We additionally assessed our best performing model using the Exact Match metric to perform qualitative and quantitative analysis.

Table 2
Kvasir-vqa answer generation results on the validation and test sets for the two approaches: Finetuned models (FT), and chain-of-thought models (CoT). Results are in (%).

Model	Meth.	Validation					Test				
		Bleu	rouge1	rouge2	rougeL	Meteor	Bleu	rouge1	rouge2	rougeL	Meteor
Qwen2-VL-7B	CoT	-	56	7	55	33	-	-	-	-	-
Qwen2-VL-32B	CoT	-	72	8	71	39	-	-	-	-	-
BLIP2-4B	CoT	-	70	12	69	39	-	-	-	-	-
Qwen2-VL-7B	FT	-	61	6	61	35	-	-	-	-	-
BLIP2-4B	FT	23	83	10	83	**46**	22	92	11	92	50

Table 3
BLIP2 wrong prediction examples for Cecum, Esophagitis and Pylorus image categories.

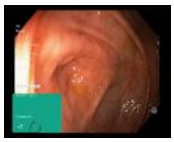
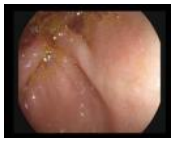
Image category	Question	Prediction	Gold answer	Image
Cecum	What color is the anatomical landmark? If more than one separate with ;	red; pink; white	`pink;red`	
Cecum	What is the size of the polyp?	5–10 mm	`none`	
Esophagitis	Are there any abnormalities in the image? Check all that are present.	ulcerative colitis	`oesophagitis`	
Esophagitis	Where in the image is the abnormality?	center; center-left; center-right	Lower-Right; Lower-Center;	
Pylorus	What type of procedure is the image taken from?	gastroscopy	`colonoscopy`	
Pylorus	What type of procedure is the image taken from?	gastroscopy	`colonoscopy`	

Table 2 show that the BLIP2 model fine-tuned on the Kvasir-VQA dataset achieved the best overall performance on the Kvsair-vqa validation set. Note that, to achieve the best performance on the test set,

Table 4
BLIP2 wrong prediction examples for Normal Colon, Normal Esophagus, Polyp and Instruments image categories.

Image category	Question	Prediction	Gold answer	Image
Normal Colon	Is there a green/black box artefact?	yes	`No`	
Normal Esophagus	Is there text?	no	`Yes`	
Polyp	What color is the abnormality? If more than one separate with ;	pink; red	`white;pink`	
Polyp	How many polyps are in the image?	0	`1`	
Instruments	Are there any instruments in the image? Check all that are present.	tube	`Polyp Snare`	
Instruments	Where in the image is the instrument?	center; lower-center; lower-right	Center; Center-Right; ...	

we further finetuned the BLIP2 model on the training sets of PathVQA [15], VQA-RAD [14], SLAKE [16] and OmniMedVQA [19] datasets allowing to achieve a METEOR score of 50 and a BLEU score of 22. In contrast, while Chain-of-Thought prompting enhances in general interpretability by providing intermediate reasoning, its practical effectiveness on the Kvasir-vqa dataset seems limited without additional rationale supervision. We believe that instruction finetuning of the Qwen2-VL-72B-instruct we used to generate the reasoning rationales on medical-domain data would help providing more comprehensive rationales (less noisy rationales) and thus alleviating the answer/rationale discrepancies.

Furthermore, the performance gap between BLIP2-Flan-T5-XL and Qwen2-VL models is worth noting. Indeed, BLIP2-Flan-T5-XL consistently outperforms Qwen2-VL-7B-instruct whatever the training method and has comparable performance with Qwen2-VL-32B-instruct in the CoT setting despite their difference in model size. Besides, given that our experiment LoRA configuration reduces the number of trainable parameters, we found that: BLIP2-Flan-T5-XL has 4.7M, Qwen2-VL-7B has 2.5M and Qwen2-VL-32 has 8.3M trainable parameters. This shows that BLIP2-Flan-T5-XL shows superior capabilities on the Kvasir-vqa task despite its relative size. We believe that encoder-decoder architectures such as BLIP2 are more suitable for VQA tasks as they allow to encode rich image features before generating the textual output, facilitating better multimodal alignment, while decoder-only models like Qwen2-VL must process the image and question together through a single stream, which may limit fine-grained control over visual and textual token interactions during generation.

Table 5 shows the exact match evaluation results of our best performing model (BLIP2-FT) by image category. We achieved the highest EM scores of 99.02% and 97.83% respectively for Normal Colon and

Table 5
Best performing BLIP2 model results by Image Category on the validation split using the Exact Match metric. Note that our best model achieved a global Exact Match score of 75.28% and F1 score of 77.22%.

Image Category	total	correct	incorrect	**Exact Match** (%)
Polyp	318	155	163	48.74
Instruments	46	28	18	60.87
Esophagitis	308	244	64	79.22
Pylorus	242	198	44	81.82
Cecum	322	271	51	84.16
Normal Esophagus	46	45	1	97.83
Normal Colon	102	101	1	99.02

Normal Esophagus image catgories. This due to the low variability in answers, as all the questions related to these image categories cover only yes/no question types which seem to be an easy task for BLIP2 model finetuned on similar data. We see in Table 4 the only examples of these image categories where our model wrongly predicts the yes/no questions. Moreover, Table 6 and Table 7 show that our model achieves respectively 96.23% and 91.94% EM scores on questions with yes/no answers whatever the image category.

Our best BLIP2 model also achieved solid EM performance on the following image categories: Cecum with 84.16%, Pylorus with 81.82%, and Esophagitis 79.22%, indicating its relative ability in identifying specific anatomical regions and whether some pathological signs are present. In contrast, our model struggled the most on questions related to the Polyp image category with the lowest EM score of 48.74%. On the one hand, answering questions about polyp requires the model to consistently identify more subtle image features and on the other hand, the polyp image category in the dataset cover a wider range of question types including among others: yes/no, color-related, counting and location-related questions. Similarly, the Instruments image category is also challenging for our model which yielded an EM score of 60.87%, as it also covers several question types which require distinguishing medical instruments from the background. These results suggest that the model may greatly benefit from more advanced and specific reasoning abilities such as visual spatial reasoning in order for example to accurately answer location-related questions for which our model achieves poor results (see Tables 6 and 7).

6. Conclusion

This paper presented two simple approaches for tackling the ImageCLEFmed 2025 MedVQA challenge using the Kvasir-VQA dataset. While the Chain-of-Thought approach offered insights into the reasoning process behind answer generation, fine-tuning large generative models achieved significantly better performance across all evaluation metrics. Our experiments demonstrate the effectiveness of large vision-language models like BLIP2 when adapted to domain-specific medical tasks. Qualitative and quantitative analysis show that endowing the model with more complex visual reasoning abilities might improve the VQA performance on the questions related to the most challenging image categories namely Polyp and Instruments.

Declaration on Generative AI

During the preparation of this work, we acknowledge the use of generative AI tools (Chat-GPT-4) for only spell checking, paraphrasing, and latex formatting purposes. After using Chat-GPT-4, we systematically reviewed and edited all 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, Sci. Data 7 (2020) 1–14. doi:10.1038/s41597-020-00622-y.

[2] 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.

[3] B. Ionescu, H. Müller, D.-C. Stanciu, A. Idrissi-Yaghir, A. Radzhabov, A. G. S. de Herrera, A. Andrei, A. Storås, A. B. Abacha, B. Bracke, B. Lecouteux, B. Stein, C. Macaire, C. M. Friedrich, C. S. Schmidt, D. Fabre, D. Schwab, D. Dimitrov, E. Esperança-Rodier, G. Constantin, H. Becker, H. Damm, H. Schäfer, I. Rodkin, I. Koychev, J. Kiesel, J. Rückert, J. Malvehy, L.-D. Ștefan, L. Bloch, M. Potthast, M. Heinrich, M. A. Riegler, M. Dogariu, N. Codella, P. H. P. Nakov, R. Brüngel, R. A. Novoa, R. J. Das, S. A. Hicks, S. Gautam, T. M. G. Pakull, V. Thambawita, V. Kovalev, W.-W. Yim, Z. Xie, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[4] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, Q. D. Tran, Overcoming data limitation in medical visual question answering, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22, Springer, 2019, pp. 522–530.

[5] L.-M. Zhan, B. Liu, L. Fan, J. Chen, X.-M. Wu, Medical visual question answering via conditional reasoning, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2345–2354.

[6] J. J. Lau, S. Gayen, A. Ben Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, Scientific data 5 (2018) 1–10.

[7] F. Ren, Y. Zhou, Cgmvqa: A new classification and generative model for medical visual question answering, IEEE Access 8 (2020) 50626–50636.

[8] B. Liu, L.-M. Zhan, L. Xu, X.-M. Wu, Medical visual question answering via conditional reasoning and contrastive learning, IEEE transactions on medical imaging 42 (2022) 1532–1545.

[9] 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.

[10] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, W. Xie, Pmc-vqa: Visual instruction tuning for medical visual question answering, arXiv preprint arXiv:2305.10415 (2023).

[11] 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, Advances in Neural Information Processing Systems 36 (2023) 28541–28564.

[12] J. Liu, Y. Wang, J. Du, J. T. Zhou, Z. Liu, Medcot: Medical chain of thought via hierarchical expert, arXiv preprint arXiv:2412.13736 (2024).

[13] X. Gai, C. Zhou, J. Liu, Y. Feng, J. Wu, Z. Liu, Medthink: Explaining medical visual question answering via multimodal decision-making rationale, arXiv preprint arXiv:2404.12372 (2024).

[14] J. J. Lau, S. Gayen, A. Ben Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, Scientific data 5 (2018) 1–10.

[15] X. He, Y. Zhang, L. Mou, E. Xing, P. Xie, Pathvqa: 30000+ questions for medical visual question answering, arXiv preprint arXiv:2003.10286 (2020).

[16] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, X.-M. Wu, Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, in: 2021 IEEE 18th international symposium on biomedical imaging (ISBI), IEEE, 2021, pp. 1650–1654.

[17] 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.
[18] M. Chaichuk, S. Gautam, S. Hicks, E. Tutubalina, Prompt to Polyp: Medical Text-Conditioned Image Synthesis with Diffusion Models, arXiv (2025). doi:10.48550/arXiv.2505.05573.
[19] Y. Hu, T. Li, Q. Lu, W. Shao, J. He, Y. Qiao, P. Luo, Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22170–22183.
[20] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: Multimodal reasoning via thought chains for science question answering, Advances in Neural Information Processing Systems 35 (2022) 2507–2521.
[21] G. Zheng, B. Yang, J. Tang, H.-Y. Zhou, S. Yang, Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models, Advances in Neural Information Processing Systems 36 (2023) 5168–5191.
[22] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, in: NIPS 2017 Workshop on Autodiff, MIT Press, Long Beach, CA, USA, 2017.
[23] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al., Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, arXiv preprint arXiv:2409.12191 (2024).
[24] 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.
[25] E. J. Hu, yelong 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, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

A. Additional Quantitative Results

The following tables shows our best performing BLIP2 model results by answer (or question type) on the validation split.

Table 6
Best performing BLIP2 model results by Question type (answer) on the validation split using the Exact Match metric.

Answer	Total	Correct	Incorrect	Exact Match (%)
2	2	0	2	0.00
3	1	0	1	0.00
5-10mm;11-20mm	1	0	1	0.00
< 5mm	3	0	3	0.00
>20mm;5-10mm	1	0	1	0.00
Biopsy Forceps	1	0	1	0.00
Cecum	3	0	3	0.00
Center	6	0	6	0.00
Center-Left	1	0	1	0.00
Center-Left;Lower-Left	2	0	2	0.00
Center;Upper-Center	1	0	1	0.00
Center-Right;Center;Upper-Center;Lower-Center;Lower-Right;Upper-Right	1	0	1	0.00
Center-Right;Lower-Center;Lower-Right	1	0	1	0.00
Center-Right;Upper-Right	1	0	1	0.00
Center;Center-Left	4	0	4	0.00
Center-Right	2	0	2	0.00
Center;Center-Left;Center-Right;Lower-Left;Lower-Center;Lower-Right	1	0	1	0.00
Center;Center-Right;Lower-Right	1	0	1	0.00
Center;Center-Right;Lower-Left;Lower-Center;Lower-Right;Upper-Right;Upper-Center;Center-Left;Upper-Left	1	0	1	0.00
Center;Center-Right;Lower-Center;Lower-Right	1	0	1	0.00
Center;Center-Left;Center-Right;Lower-Left;Lower-Center	1	0	1	0.00
Center;Center-Right	2	0	2	0.00
Center;Center-Left;Upper-Right	1	0	1	0.00
Center;Center-Right;Lower-Center	2	0	2	0.00
Center;Center-Left;Upper-Center;Lower-Center	1	0	1	0.00
Center;Center-Left;Upper-Center	1	0	1	0.00
Center;Center-Left;Center-Right;Upper-Right;Upper-Center;Lower-Center;Lower-Left;Lower-Right;Upper-Left	1	0	1	0.00
Center;Center-Left;Center-Right;Lower-Left;Lower-Center;Lower-Right;Upper-Right;Upper-Center;Upper-Left	1	0	1	0.00
Center;Center-Left;Upper-Left	1	0	1	0.00
pink;red;white	9	0	9	0.00
pink	25	1	24	4.00
11-20mm	7	1	6	14.29
5-10mm	17	4	13	23.53
1	35	16	19	45.71
Polyp	35	20	15	57.14
>20mm	5	3	2	60.00
gastroscopy	22	15	7	68.18
colonoscopy	51	35	16	68.63
Paris Ip	10	7	3	70.00
oesophagitis	19	16	3	84.21
No	211	194	17	91.94
0	146	138	8	94.52
none	292	277	15	94.86
Yes	212	204	8	96.23
not relevant	73	72	1	98.63
Oesophagitis	2	2	0	100.00
Not relevant	1	1	0	100.00
Colonoscopy	36	36	0	100.00

Table 7
Best performing BLIP2 model results by Question type (answer) on the validation split using the Exact Match metric.

Answer	Total	Correct	Incorrect	Exact Match (%)
yellow	4	0	4	0.00
pink;red;purple	2	0	2	0.00
Lower-Center;Lower-Right;Center;Center-Right	1	0	1	0.00
Lower-Right	1	0	1	0.00
Lower-Right;Lower-Center;Lower-Left;Center-Right;Center;Center-Left;Upper-Right;Upper-Center;Upper-Left	5	0	5	0.00
white;red	3	0	3	0.00
white	1	0	1	0.00
red;pink	2	0	2	0.00
Paris IIa	6	0	6	0.00
Paris Ip;Paris Is	1	0	1	0.00
Paris Is	18	0	18	0.00
Polyp Snare	3	0	3	0.00
Ulcerative Colitis	1	0	1	0.00
Upper-Center	1	0	1	0.00
Upper-Center;Center	1	0	1	0.00
Upper-Center;Center-Left;Center;Lower-Left;Lower-Center	1	0	1	0.00
Upper-Center;Upper-Left;Center-Left;Center;Lower-Center	1	0	1	0.00
Upper-Center;Upper-Left;Upper-Right;Center;Center-Left;Center-Right;Lower-Center;Lower-Right;Lower-Left	1	0	1	0.00
pink;red	36	0	36	0.00
pink/red	3	0	3	0.00
pink;red;yellow	17	0	17	0.00
pink;white	2	0	2	0.00
red	2	0	2	0.00
Upper-Left;Upper-Center;Upper-Right;Center-Left;Center;Center-Right;Lower-Left;Lower-Center;Lower-Right	1	0	1	0.00
Lower-Center	1	0	1	0.00
Upper-Left;Upper-Center;Center-Left;Center;Upper-Right;Center-Right;Lower-Left;Lower-Center;Lower-Right	1	0	1	0.00
Upper-Left;Upper-Center;Center-Left;Center	1	0	1	0.00
Upper-Left;Upper-Center;Center-Left	1	0	1	0.00
Upper-Left;Upper-Center	1	0	1	0.00
Upper-Left;Center-Left	1	0	1	0.00
Upper-Left	2	0	2	0.00
Upper-Center;Upper-Right	1	0	1	0.00
Upper-Left;Upper-Center;Center-Left;Center;Lower-Left;Lower-Center	1	0	1	0.00
white;pink	2	0	2	0.00
Lower-Center;Lower-Right	1	0	1	0.00
Center;Lower-Center	3	0	3	0.00