# Multi-Task Learning for Visually Grounded Reasoning in Gastrointestinal VQA

Itbaan Safwan[1], Muhammad Annas Shaikh[1], Muhammad Haaris[1], Ramail Khan[1] and Muhammad Atif Tahir[1]

[1]*School of Mathematics and Computer Science, Institute of Business Administration (IBA), Karachi, Pakistan*

Abstract

We present a multi-task framework for the MediaEval Medico 2025 challenge, leveraging a LoRA-tuned Florence-2 model for simultaneous visual question answering (VQA), explanation generation, and visual grounding. The proposed system integrates three curated datasets: (1) Kvasir-VQA-x1 for question-answer learning, (2) a synthetically enriched explanation dataset offering structured medical reasoning, and (3) text-to-region pairs linking visual features with segmentation masks. This multi-task setup enables the model to jointly learn visual grounding, reasoning, and interpretation, producing responses that are both accurate and interpretable. Extensive evaluation demonstrates that our approach substantially improves over single-task baselines in both answer accuracy and visual localization, highlighting the effectiveness of grounded multi-task learning for medical VQA applications. Our code: `GitHub Repository`

## 1. Introduction

The MediaEval Medico 2025 VQA Challenge [1] focuses on explainable AI for Gastrointestinal imaging. It features two subtasks—visual question answering and multimodal explanation generation—based on the Kvasir-VQA-x1 dataset [2].

Building on this challenge, we address a key limitation of Vision-Language Models (VLMs): while VLMs recognize visual features well, they often lack medical terminology and reasoning for trustworthy clinical applications. Standard VQA training without explicit visual grounding or detailed explanations is insufficient for medical AI. We propose a multi-task learning framework enriching training with textual explanations and visual grounding through two curated datasets: (1) textual explanations providing medical reasoning for complexity-1 questions, and (2) visually grounded data linking answers to segmentation masks. Training Florence-2 [3] with this framework enables simultaneous learning of visual grounding, medical reasoning, and question answering, improving VQA performance while producing models grounded in visual evidence and medical knowledge.

The principles guiding our framework draw from prior work that has consistently shown multi-task learning to be effective for vision-language models. [4] trained on 12 tasks, reducing parameters from 3B to 270M while improving performance by 2.05 points. Molmo [5] showed that curating detailed captioning, QA pairs, and 2D pointing enables smaller models to outperform larger closed-source models. In medical imaging, [6] achieved Dice score gains of 0.28%-14.47% using multi-modal pre-training on 2.4M brain MRI images. We apply these principles to medical VQA, combining visual grounding, explanation generation, and question answering for GI image analysis.

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

i.safwan.26197@khi.iba.edu.pk (I. Safwan); m.shaikh.26919@khi.iba.edu.pk (M. A. Shaikh); m.haaris.27083@khi.iba.edu.pk (M. Haaris); r.khan.26924@khi.iba.edu.pk (R. Khan); atiftahir@iba.edu.pk (M. A. Tahir)

CEUR Workshop Proceedings (CEUR-WS.org)

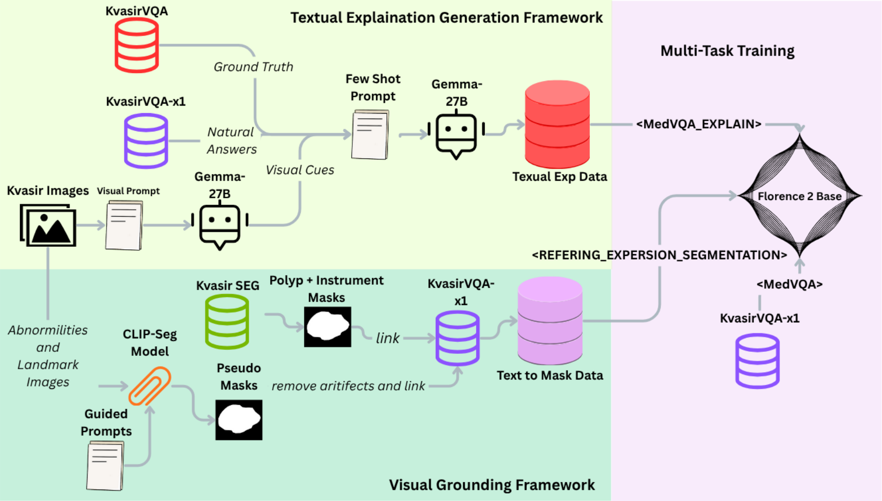


**Figure 1:** Our proposed multi-task training framework.

# 2. Methodology

## 2.1. Textual and Visual Generation Framework

**Textual Explanation Dataset** was curated in two stages. First, starting with simple image-QA pairs from [2], we enriched them with ground-truth metadata (e.g., abnormality type, location) from [7]. We then used the Gemma-27B API [8] to generate further descriptive visual cues. The prompting strategy for Gemma explicitly discouraged medical terminology, focusing instead on observable features (shape, texture) to improve generalization. In the second stage, we used the Gemma-27B model with few-shot prompting [9] to synthesize the ground-truth metadata, visual descriptions, and original QA pairs into complete, well-structured textual explanations. After post-processing to enhance fluency, the final dataset contained approximately 3,344 samples.

**Text-to-Relevant-Region Dataset** was curated by first collecting segmentation masks. We generated pseudo-masks for abnormalities and landmarks (2,954 samples) using the ClipSeg model [10] with descriptive text prompts (e.g., "red patches"). To improve coverage, we used multiple prompts per case and refined the resulting masks by removing black backgrounds and artifacts using OpenCV. We also incorporated 1,383 high-quality polyp and instrument masks from the Kvasir-SEG datasets [11, 12]. In the second stage, we linked these masks to corresponding textual answers from [2]. The final dataset comprises 4,337 text–mask pairs. Although pseudo-masks lack medical precision, they offer useful approximate supervision, guiding the model to ground predictions in relevant visual evidence.

## 2.2. Training and Post-processing

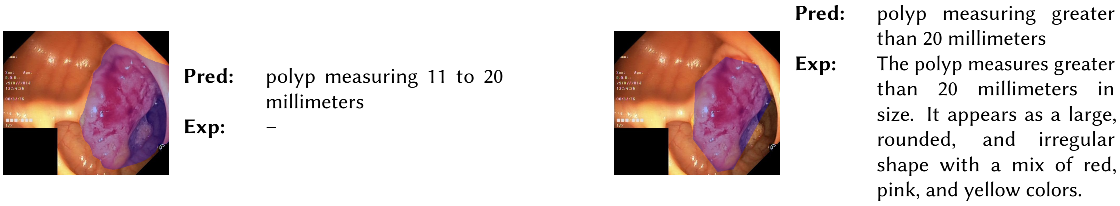
The Florence-2 model was fine-tuned using LoRA on a combined dataset (Kvasir-VQA-x1, textual explanation, text-to-region) with an 80/20 split to prevent image overlap between training and validation sets. Training was conducted on Kaggle ($2\times$T4 GPUs) for one epoch using the Transformers Trainer library. We applied LoRA (rank=128, alpha=256) to all attention modules and trained with a learning rate of $5 \times 10^{-5}$, warmup_ratio=0.1, FP16 precision, and an effective batch size of 12 (2 per device × 3 accumulation steps).

For post-processing and inference, we used three task-specific tokens: `<MedVQA>` for VQA, `<MedVQA_EXPLAIN>` for explanations, and `<REFERRING_EXPRESSION_SEGMENTATION>` for region grounding (masks were converted to Florence-2 location tokens). **Subtask 1** generated outputs using `<MedVQA>` with the question as input. **Subtask 2** followed a two-step approach: first predicting the answer with `<MedVQA>`, then localizing the region using `<REFERRING_EXPRESSION_SEGMENTATION>` with the predicted answer. Explanations were generated using `<MedVQA_EXPLAIN>` with the question and the prompt "Explain in detail". A confidence score for each generated explanation was derived from decoding stability, computed as the average top-$k$ ($k = 5$) probability mass over top tokens. We found that conflicting explanations scored lower, while correct ones scored higher, though further validation is needed.

**Table 1**
Validation Scores Across Different Model Variants. FL2: Florence-2; MT: Multi-task Training; Format: Rank_Alpha (e.g., 32_64 indicates LoRA rank=32, alpha=64)

| Model Configuration | Seg-IoU | Seg-IoU | Seg-IoU | BLEU | ROUGE-1 | ROUGE-2 | ROUGE-L | METEOR |
|---|---|---|---|---|---|---|---|---|
| | Instrument | Polyp | Pseudo | | | | | |
| FL2_VQA_32_64 | 0.4911 | 0.4029 | 0.1309 | 0.4348 | 0.6856 | 0.4935 | 0.6566 | 0.6548 |
| FL2_VQA_64_128 | 0.3712 | 0.5589 | 0.1651 | 0.4518 | 0.6953 | 0.5074 | 0.6675 | 0.6674 |
| FL2_VQA_128_256 | 0.2961 | 0.2390 | 0.1688 | 0.4623 | 0.7024 | 0.5163 | 0.6754 | 0.6735 |
| FL2_VQA_MT_32_64 | 0.7098 | 0.6828 | 0.5110 | 0.4432 | 0.6894 | 0.5010 | 0.6613 | 0.6592 |
| FL2_VQA_MT_64_128 | 0.7374 | **0.7063** | 0.5344 | 0.4615 | 0.7008 | 0.5144 | 0.6731 | 0.6728 |
| FL2_VQA_MT_128_256 | **0.7403** | 0.6879 | **0.5447** | **0.4726** | **0.7076** | **0.5234** | **0.6802** | **0.6804** |
| **Official results on Private Dataset (FL2_VQA_MT_128_256):** | | | | | | | | |
| FL2_VQA_MT_128_256 | – | – | – | 0.4539 | 0.6828 | 0.4954 | 0.6531 | 0.6515 |

Additional results on private dataset: CHRF++ = 63.79, BERTScore F1 = 0.9479.



**Pred:** polyp measuring 11 to 20 millimeters
**Exp:** –

**Pred:** polyp measuring greater than 20 millimeters
**Exp:** The polyp measures greater than 20 millimeters in size. It appears as a large, rounded, and irregular shape with a mix of red, pink, and yellow colors.

**Figure 2:** Comparison of model responses before and after multi-task training. **Question:** What is the size of the polyp? **Actual Answer:** polyp larger than 20 millimeters.

## 3. Results and Evaluation

Increasing LoRA parameters, as referenced in Table 1, reveals a critical trade-off in VQA-only training. While language scores improved, with BLEU increasing from 0.4348 to 0.4623, visual grounding severely degraded as Seg-IoU Instrument scores dropped from 0.4911 to 0.2961. This suggests the model overfits to linguistic cues. Conversely, our multi-task approach improved all metrics simultaneously; Seg-IoU Instrument scores rose from 0.7098 to 0.7403 and BLEU scores climbed from 0.4432 to 0.4726. This culminated in our best-performing model (FL2_VQA_MT_128_256), which demonstrated robust generalization to unseen data on the official private dataset with a BLEU of 0.4539, ROUGE-L of 0.6531, and BERTScore F1 of 0.9479. This shows the segmentation task acts as an effective regularizer, preserving visual understanding while enhancing language quality. This is evident in Figure 2, where the multi-task model produces a more accurate polyp segmentation and correctly identifies it as "greater than 20 millimeters," consistent with the ground truth.

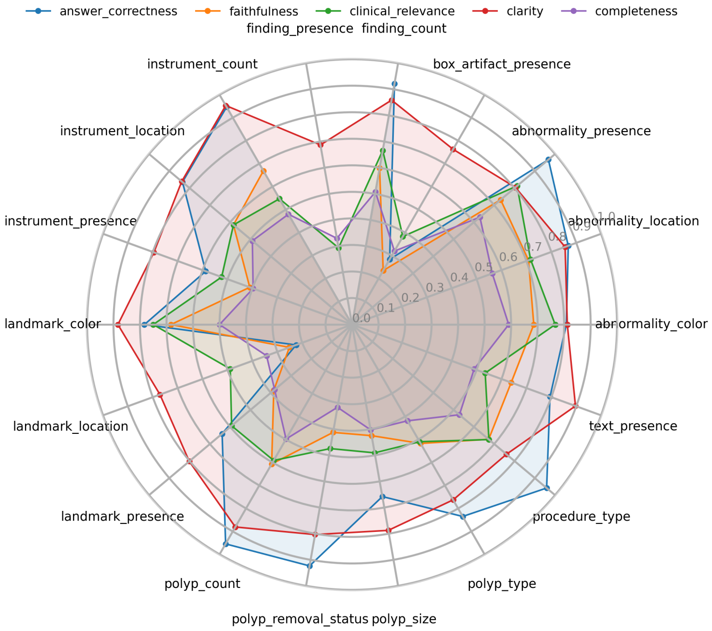


**Figure 3:** Sub-Task 2: Question-wise radar graph for each explainability metric on official results.

As shown in the radar chart (Figure 3), our model's textual explanations for Sub-Task 2 are a mixed success. For structured tasks like presence detection and counting, the model achieves high **answer correctness** ($> 0.9$) and **clarity** ($> 0.8$). However, it shows significant weaknesses in **completeness** (0.3–0.7) and **faithfulness** (0.4–0.6). The model excels with binary and quantitative attributes but struggles with descriptive ones, particularly color identification, where all metrics fall below 0.6. These results highlight that although our training has improved correctness, further refinement is essential to ensure explanations are complete and clinically reliable.

## 4. Conclusion and Future Work

Our work advances grounded training for VLMs in GI endoscopy but revealed key challenges. Synthetic explaination data generation using the Gemma-27B model produced less diverse descriptions, hindering effective training. Furthermore, severe class imbalance hindered visual grounding, as rare landmarks were overshadowed by prevalent polyps, and the lack of negative examples in the augmented datasets may have also biased the model training. Future work could leverage the model's pseudo-masking to propose regions for refinement, using them as prompts for the Segment Anything Model (SAM) [13], which can generate more precise masks that may be used to train an auxiliary segmentation head alongside VQA training. Additional improvements can be made by dataset augmentation for rare classes, incorporating negative examples, and grounding explanations by expert descriptions.

## Declaration on Generative AI

Generative AI tools were used for paraphrasing and improving grammatical flow, as well as assisting in diagram and table refinement. All content was carefully verified by the authors.

## References

[1] 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.

[2] 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.

[3] B. Xu, Z. Guo, A. Liu, J. Li, L. Zhang, X. Zhang, C. Wang, Florence-2: Advancing a unified representation for a variety of vision tasks, ArXiv e-prints (2024). arXiv:2311.06242.

[4] J. Lu, C. Clark, R. Zellers, R. Mottaghi, A. Kembhavi, 12-in-1: Multi-task vision and language representation learning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 10437–10446. doi:10.1109/CVPR42600.2020.01045.

[5] M. Deitke, C. Clark, Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL: https://arxiv.org/abs/2409.17146. arXiv:2409.17146.

[6] S. Rui, L. Chen, Z. Tang, L. Wang, M. Liu, S. Zhang, X. Wang, Multi-modal vision pre-training for medical image analysis, 2025. URL: https://arxiv.org/abs/2410.10604. arXiv:2410.10604.

[7] 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.

[8] Gemma Team, Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024). doi:10.48550/arXiv.2403.08295. arXiv:2403.08295, the Gemma-27B variant builds on this foundational work.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

[10] T. Lüddecke, A. S. Ecker, Image segmentation using text and image prompts, 2022. URL: https://arxiv.org/abs/2112.10003. arXiv:2112.10003.

[11] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. de Lange, P. Halvorsen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: International Conference on Multimedia Modeling, Springer, 2020, pp. 451–462. doi:10.1007/978-3-030-37734-2_37.

[12] 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, D. Johansen, P. Halvorsen, Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, in: MultiMedia Modeling, Springer International Publishing, Cham, 2021, pp. 218–229.

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything, 2023. URL: https://arxiv.org/abs/2304.02643. arXiv:2304.02643.