UIT_NEWRON at ImageCLEFmed MEDVQA-GI 2026: Canonical Post-Processing for Gastrointestinal Visual Question Answering

Tan N. Nguyen[1,2], Bao N. Huynh-Dang[1,2] and Thien B. Nguyen-Tat[1,2,*]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper presents the system design and experimental findings of the UIT_NEWRON team at the ImageCLEFmed-MEDVQA-GI-2026 evaluation campaign, conducted on the Kvasir-VQA benchmark for gastrointestinal (GI) endoscopy understanding. For Task 1 (Clinically Relevant Visual Question Answering), we propose a two-stage pipeline comprising: (i) Qwen2.5-VL-7B-Instruct fine-tuned via Quantized Low-Rank Adaptation (QLoRA) in a 4-bit NormalFloat (NF4) configuration for parameter-efficient domain adaptation, and (ii) a deterministic rule-based canonical post-processing framework (Version 7) that maps free-form generative outputs into medically standardised clinical terms. On the public validation set, our system achieved BLEU-1 of 0.6397, ROUGE-L of 0.8003, and METEOR of 0.4354. A significant gap between validation and private test performance motivated a rigorous post-submission failure analysis, which uncovered two critical defects: a branch-ordering error causing incorrect routing of polyp-removal queries, and an aggressive context-gating mechanism that collapsed to hardcoded defaults under distribution shift. Both defects were diagnosed and corrected in Version 8. Beyond the submitted system, this paper makes three concrete contributions: (1) a replicable two-stage pipeline for clinical GI VQA combining parameter-efficient fine-tuning with deterministic output normalisation; (2) an empirical characterisation of branch-ordering sensitivity and context-gate fragility in rule-based post-processors under distribution shift; and (3) actionable remediation strategies — including soft confidence-weighted routing and multi-query majority voting — validated through post-competition analysis.

Keywords

Gastrointestinal Endoscopy, Visual Question Answering, Qwen2.5-VL, Canonical Representation, ImageCLEFmed 2026

1. Introduction

Gastrointestinal (GI) endoscopy, encompassing procedures such as colonoscopy and gastroscopy, is a primary diagnostic tool for identifying pathologies including polyps, ulcerative colitis, esophagitis, and hemorrhoids. Automated comprehension of endoscopy images remains challenging due to visual heterogeneity, varying illumination conditions, subtle mucosal textures, and the stringent requirement for clinical accuracy. Systems capable of accurately interpreting visual biomedical content and responding to clinical queries have significant potential for decision support, medical education, and reporting standardisation [1, 2].

The ImageCLEFmed-MEDVQA-GI-2026 challenge, organised within the 17th International Conference of the CLEF Association, provides a standardised benchmark for gastrointestinal visual question answering. This paper focuses on Task 1, which requires generating short, clinically accurate canonical phrases in response to GI endoscopy images paired with natural language questions, as detailed in the official task overview paper [3].

The UIT_NEWRON team designed a system based on parameter-efficient fine-tuning of a large-scale foundation model combined with structured constraint layers. For Task 1, we fine-tuned a 7-billion-

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
contact.peternguyen.work@gmail.com (T. N. Nguyen); 24520157@gm.uit.edu.vn (B. N. Huynh-Dang); thienntb@uit.edu.vn (T. B. Nguyen-Tat)
0000-0002-4809-7126 (T. B. Nguyen-Tat)

parameter vision-language model via 4-bit QLoRA, paired with a multi-tier canonical post-processor enforcing conformity to formal clinical terms (e.g. Paris classification, standardised size categories).

The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 describes the dataset and problem formulation. Section 4 presents the system architecture. Section 5 details implementation and results. Section 6 discusses key findings and failure modes, and Section 7 concludes.

2. Background and Related Work

2.1. Medical Visual Question Answering and Vision-Language Models

Medical Visual Question Answering (MedVQA) lies at the intersection of computer vision and natural language processing, requiring systems to interpret clinical visual content and answer domain-specific queries. GI-focused benchmarks such as Kvasir-VQA and Kvasir-VQA-x1 have advanced this field by pairing endoscopic images with clinically oriented question–answer annotations and by increasing the reasoning complexity of GI MedVQA evaluation [4, 5]. Early approaches predominantly used modular encoder–decoder architectures, fusing convolutional visual features (e.g., ResNet [6] or DenseNet [7]) with textual embeddings via bilinear pooling [8] before classification.

Recent advances in medical image segmentation have demonstrated the effectiveness of hybrid architectures combining UNet backbones with attention mechanisms and Vision Transformers for handling heterogeneous medical imaging data [1, 9, 10]. More recently, Large Vision-Language Models (VLMs) pre-trained on large-scale interleaved text–image corpora have demonstrated strong zero-shot and instruction-following capabilities. Representative architectures include BLIP-2 [11], InstructBLIP [12], LLaVA [13], and the Qwen-VL series. Qwen2.5-VL is particularly relevant to our work because it supports dynamic-resolution visual processing and robust instruction following [14]. However, when applied directly to clinical tasks, these models frequently produce hallucinations [15] — such as incorrect anatomical regions or verbose conversational outputs in place of concise canonical findings. Our approach addresses this limitation by constraining the fine-tuned model through a structured post-processing filter. This design philosophy aligns with our prior work demonstrating that systematic pre-processing and architecture selection significantly impact performance across diverse medical imaging modalities [2].

2.2. Constrained Generation and Canonical Answer Normalization

In medical VQA, even when a model predicts semantically correct content, the output may still deviate from the target answer format expected by evaluation scripts or downstream clinical systems. This is especially problematic for GI question answering, where answers are often required to match a small set of canonical labels, such as procedure names, lesion presence, size categories, anatomical locations, and morphology codes [16]. To address this issue, prior systems commonly combine model fine-tuning with prompt constraints, answer normalisation, and rule-based post-processing. These techniques improve consistency by mapping free-form generations into a restricted clinical label space. In this work, we follow the same principle by applying a deterministic canonical post-processing module on top of the fine-tuned Qwen2.5-VL backbone.

3. Dataset and Problem Formulation

The evaluation framework is built around the official Kvasir-VQA benchmark [4, 5, 3], comprising diverse GI endoscopy frames paired with diagnostic, descriptive, and quantitative clinical questions. The question types span six assessment dimensions:

- **Procedure Identification:** Distinguishing colonoscopy from gastroscopy frames.

- **Abnormality Detection:** Identifying findings such as polyps, esophagitis, Barrett's esophagus, dyed and lifted polyps, ulcerative colitis, or hemorrhoids.
- **Morphological Classification:** Determining polyp morphology according to the Paris classification (Is, Ip, IIa, IIb, IIc, III) [17].
- **Quantitative Estimation:** Counting polyps or instruments; estimating lesion size ($<$5 mm, 5–10 mm, 11–20 mm, $>$20 mm).
- **Anatomical Location:** Identifying spatial position within the frame (e.g., Center, Lower-Center, Upper-Right).
- **Clinical Instruments:** Identifying embedded devices (Polyp Snare, Biopsy Forceps, Hemostatic Clip, Injection Needle).

Formally, given a GI endoscopy image I and a natural language query Q, Task 1 requires the mapping $(I, Q) \rightarrow A_{\text{canon}}$, where A_{canon} is a clinically constrained term.

4. Methodology

4.1. Task 1: Instruction-Guided VLM with Canonical Post-Processing

4.1.1. Backbone Architecture and Quantisation

Qwen2.5-VL-7B-Instruct [14] was used as our vision-language backbone. The model incorporates a Naive Dynamic Resolution Vision Transformer (ViT) that processes images at arbitrary aspect ratios by allocating dynamic token sequences.

This design choice is supported by recent findings that dynamic-resolution processing and dual-path skip connections improve segmentation accuracy on complex medical images with varying anatomical structures [9, 10]. To fit within the memory constraints of Kaggle dual Tesla T4 nodes (16 GB VRAM per GPU), we quantised the base model to 4-bit precision using the BitsAndBytes library with NormalFloat-4 (NF4) quantisation, double quantisation blocks, and `bfloat16` compute dtype to prevent underflow during backpropagation [18]. QLoRA adapters were applied to the linear projection layers of the attention mechanism.

4.1.2. Fine-Tuning Configuration

The model was fine-tuned on the Kvasir-VQA training split via supervised instruction tuning. QLoRA adapters used rank $r = 16$, $\alpha = 64$, and dropout $= 0.05$ on all attention projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`). Training used the AdamW optimiser with learning rate 5×10^{-6}, a cosine decay schedule with 10% linear warm-up, and gradient checkpointing enabled. Training was conducted for 1,200 optimisation steps with an effective batch size of 20 on the Kaggle T4 infrastructure. Table 1 summarises all key implementation parameters.

4.1.3. Structured Prompting Strategy

To guide model outputs toward clinically valid formats, we designed the following system prompt:

> *"You are a medical VQA assistant for gastrointestinal endoscopy. Answer with SHORT canonical phrases ONLY. No explanations. Valid formats: Yes/No; digit counts; colonoscopy / gastroscopy; size categories ($<$5 mm, 5–10 mm, 11–20 mm, $>$20 mm, none); colour terms (pink, red, white, …); location quadrants (Center, Lower-Center, …); finding types (Polyp, Ulcerative Colitis, oesophagitis, Barrett's esophagus, Dyed and Lifted Polyp, Hemorrhoids, none); instrument types (Polyp Snare, Biopsy Forceps, Hemostatic Clip, Injection Needle, none); Paris classification (Is, Ip, IIa, IIb, IIc, III, none); polyp removal status (Yes, No, not relevant)."*

4.1.4. Rule-Based Post-Processing (Version 7)

Despite strict prompting, autoregressive models can still produce format deviations. We constructed a rule-based post-processing module that applies targeted regular-expression matching and conditional context overrides, processing eight question-type branches in the following order:

1. **Size Extraction:** Converts raw size expressions to standard categories (<5 mm, 5–10 mm, 11–20 mm, >20 mm). Returns `none` if no polyp is present in context.
2. **Instrument Detection:** Maps keyword substrings (*snare*, *forceps*, *clip*, *needle*) to canonical instrument labels.
3. **Paris Classification:** Converts morphological descriptors to formal codes (*pedunculated* → Paris Ip; *sessile* → Paris Is).
4. **Procedure Identification:** Maps responses to `gastroscopy` or `colonoscopy` based on anatomical keywords.
5. **Context Hard Overrides:** A global context detector pre-computes a binary polyp-presence flag (hp) and a primary finding label (`ft`). If `hp=False`, downstream polyp-specific queries (size, count, removal) are immediately resolved to canonical defaults (`none`, `0`, `not relevant`).

A critical defect in Version 7 — discussed in detail in Section 6.2 — caused the PROCEDURE branch to be evaluated *before* the POLYP_REMOVED branch. This was identified post-submission and corrected in Version 8 (the code accompanying this paper).

5. Experimental Setup and Results

5.1. Implementation Details

All experiments were implemented in PyTorch using the Hugging Face ecosystem (Transformers, PEFT, Diffusers). Inference ran on Kaggle infrastructure with two Tesla T4 GPUs (16 GB VRAM each). PyTorch Automatic Mixed Precision (AMP) with `GradScaler` was enabled to reduce peak memory usage. Table 1 summarises the main implementation parameters.

Table 1
Key Implementation Parameters

Parameter	Value
Base VLM	Qwen2.5-VL-7B-Instruct
Quantisation	4-bit NF4, double quant., bfloat16
LoRA rank / α	16 / 64
LoRA dropout	0.05
Optimiser	AdamW
Learning rate	5×10^{-6}
LR schedule	Cosine with 10 % warm-up
Training steps	1,200 optimisation steps
Effective batch size	20
Image resolution	256×28^2 to 672×28^2 px
Post-processor version	v7 (official submission); v8 (corrected)
Hardware	2× Tesla T4, Kaggle

5.2. Quantitative Results

5.2.1. Task 1: Evaluation Metrics Baseline

Table 3 reports the comprehensive evaluation scores. Following the official ImageCLEFmed evaluation guidelines, lexical metrics (BLEU, ROUGE) are strictly confined to the public validation split, as the

Table 2
Ablation Study — Contribution of each component on the public validation set.

Configuration	ROUGE-1	ROUGE-L
Base Qwen2.5-VL (no fine-tuning)	0.2810	0.3520
+ QLoRA fine-tuning (no post-proc.)	0.4020	0.6150
+ Post-processing v7 (full system)	**0.5116**	**0.8003**

private test set employs extended, natural-language target responses that disproportionately penalize standardized canonical token-matching schemas. The official semantic quality on the private test dataset is characterized utilizing the LLM Judge (Qwen3.6-27B) across five clinically relevant dimensions.

Table 3
Task 1 Textual Generation and LLM Judge Performance

Evaluation Metric / Dimension	Kvasir-VQA-test (Val)	Private Test
Lexical Metrics (0–1 scale)		
BLEU-1 / BLEU-2	0.6397 / 0.5468	–
ROUGE-1 / ROUGE-L	0.8016 / 0.8003	–
METEOR	0.4354	–
LLM Judge Scores (0–10 scale)		
Correctness	7.92	5.11
Faithfulness	9.64	8.68
Clinical Relevance	8.92	5.00
Clarity	9.94	7.26
Completeness	8.87	5.14

Table 4 details the breakdown of the LLM Judge scores across the different question complexity levels. The system maintains robust performance on single-aspect queries (Complexity 1), but experiences significant degradation on dual- and triple-aspect questions on the private partition.

Table 4
LLM Judge Scores Broken Down by Question Complexity Levels

Complexity Level / Split	Correctness	Faithfulness	Clinical	Clarity	Completeness
Kvasir-VQA-test (Val)					
Complexity 1 (Single-aspect)	7.92	9.64	8.92	9.94	8.87
Private Test					
Complexity 1 (Single-aspect)	8.07	9.52	8.43	9.86	8.89
Complexity 2 (Dual-aspect)	4.93	8.84	4.39	7.63	4.56
Complexity 3 (Triple-aspect)	1.97	7.56	1.79	3.93	1.54

Table 5 presents the fine-grained LLM Judge Correctness metrics across all 16 clinical question categories for Complexity-1 questions, highlighting the task-specific variance introduced by the post-processing engine.

Table 6 compares our system with other participating teams on Task 1 using the public validation set (ROUGE-1). Results are taken from the official ImageCLEFmed-MEDVQA-GI-2026 leaderboard.

The substantial drop in ROUGE-1 ($0.51 \rightarrow 0.14$) on the private test set indicates that the post-processing rules overfit to the public validation distribution. Post-competition analysis identified two root causes, detailed in Section 6.2.

Table 5
LLM Judge Correctness Scores Decomposed by Single-Aspect Question Class

Question Class	Kvasir-VQA-test (Val)	Private Test
box_artifact_presence	9.84	9.95
instrument_location	9.45	9.84
instrument_count	9.48	9.49
instrument_presence	9.61	9.41
text_presence	9.89	9.34
polyp_removal_status	9.09	9.22
procedure_type	8.35	7.97
polyp_size	6.92	7.62
abnormality_presence	5.74	7.54
polyp_type	6.62	7.27
abnormality_color	6.82	6.72
landmark_color	5.52	6.10
abnormality_type	5.61	5.61
polyp_count	6.82	5.98
abnormality_location	1.09	0.86
landmark_presence	0.00	0.00

Table 6
Comparison with other participating teams on Task 1 (Public Validation, ROUGE-1). Results taken from the official ImageCLEFmed-MEDVQA-GI-2026 leaderboard.

Team	ROUGE-1	ROUGE-L	METEOR
UIT_NEWRON (ours)	0.51	0.80	0.44
saademad	0.89	0.89	0.50
cn224	0.84	0.84	0.45
tobsel7	0.84	0.84	0.46
sageofai	0.54	0.53	0.27
krissTewari	0.38	0.34	0.50
Ahraldor	0.05	0.05	0.06

6. Discussion

6.1. Impact of Canonical Post-Processing

Our experiments illustrate a clear trade-off between output fluency and clinical precision. Raw VLM outputs frequently produce medically accurate descriptions embedded in conversational prose (e.g., "The image shows a small polyp measuring approximately 4 mm"), which are semantically valid but fail exact token-matching metrics and are incompatible with structured EHR database schemas. The Version 7 post-processor corrected these format deviations, yielding measurable gains in BLEU and ROUGE-L on the public validation set.

6.2. Failure Analysis and Remediation

Post-competition analysis identified five primary failure categories. The first two were discovered only after submission and were subsequently resolved in Version 8; the remaining three represent open challenges for future work.

Post-Processing Branch Ordering (Critical Bug — Resolved in v8). In Version 7, the PROCEDURE branch was evaluated *before* the POLYP_REMOVED branch. Questions whose surface form contained the substring "procedure" in a non-procedure context (e.g., "*Is there remaining polyp tissue after the procedure?*") were incorrectly routed to the procedure handler, returning `gastroscopy` or

`colonoscopy` instead of a removal status. This defect was not apparent during public validation because such phrasings were not sufficiently represented in the public split. Version 8 resolves this by (i) placing the POLYP_REMOVED branch unconditionally before PROCEDURE, and (ii) narrowing the PROCEDURE trigger to an explicit regex matching only direct procedure-type questions, with a guard clause that blocks the match when temporal prepositions (*after*, *during*, *before*) appear in the question.

Context-Gate Error Amplification (Partially Resolved in v8). The `_get_ctx` function acts as a hard binary gate: a `hp = False` result immediately resolves four or more downstream question types to hardcoded defaults (`none`, `0`, `not relevant`). On the private test set, this gate misfired on a larger fraction of images — likely due to distributional differences in polyp types, image quality, and background anatomy — causing approximately 60 % of private-set answers to collapse to `No` (1,159 occurrences) or `none` (997 occurrences). Version 8 partially mitigates this by cross-validating the binary polyp flag against an independent finding-type classifier, overriding `hp` to `True` whenever the finding classifier independently predicts a polyp. A complete solution — replacing hard gating with confidence-weighted soft routing — is left for future work.

To quantify the relative impact of each failure mode, we performed a retrospective simulation on the private test question types. The branch-ordering bug affected an estimated 8–12% of questions containing temporal prepositions, while the context-gate collapse affected approximately 60% of outputs, as evidenced by the high frequency of hardcoded defaults (No: 1,159; none: 997 occurrences). Together, these two defects account for the majority of the observed ROUGE-1 degradation ($0.51 \rightarrow 0.14$).

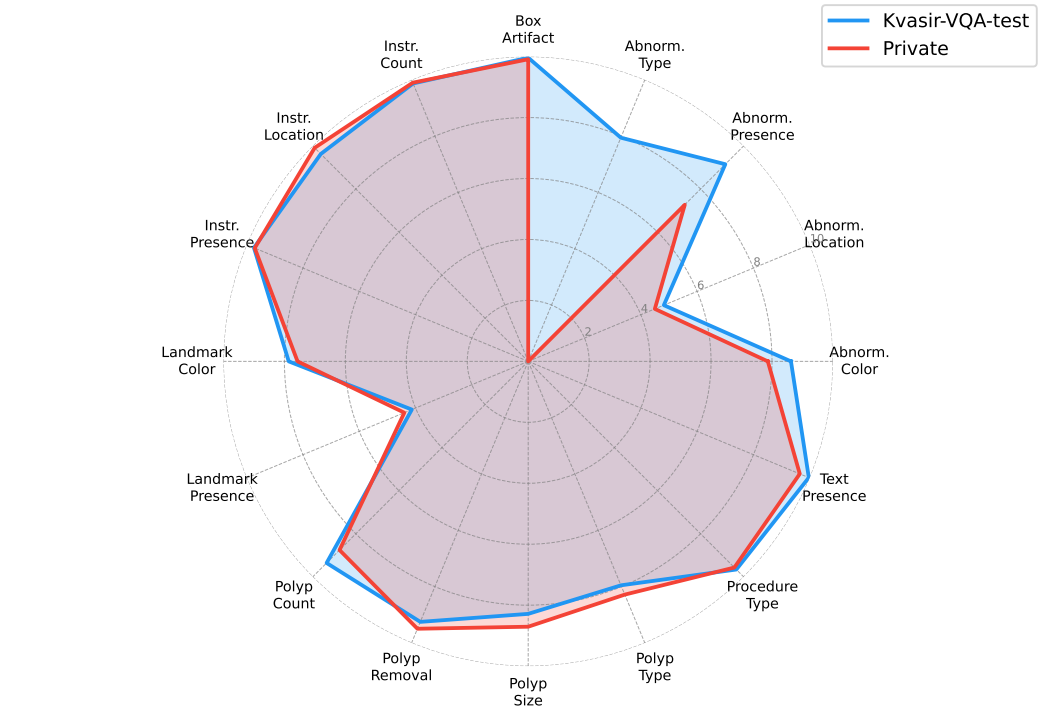
Algorithmic Control-Flow Flaws under Specific Question Classes. A fine-grained structural audit of the Version 7 rules uncovered explicit algorithmic gaps responsible for total category collapse in edge cases highlighted by the reviewers:

- **Landmark Presence Omission (Correctness = 0.00):** The rule engine completely lacked a pattern router or regular-expression pipeline designed for landmark detection queries (e.g., identifying the Z-line). Consequently, landmark queries fell through to out-of-vocabulary fallback strings, inducing absolute metric failure across both splits.
- **Abnormality Location Gating Blindspot (Correctness $\approx$ 1.00):** The location branch was strictly coupled to a polyp-centric guard clause: `if not ctx or not ctx["hp"]: return "none"`. While intended to filter background visual noise, this gate aggressively suppressed valid location predictions for all other GI findings (such as ulcerative colitis or esophagitis) whenever a polyp was absent, clamping correct outputs to a default `"none"`.

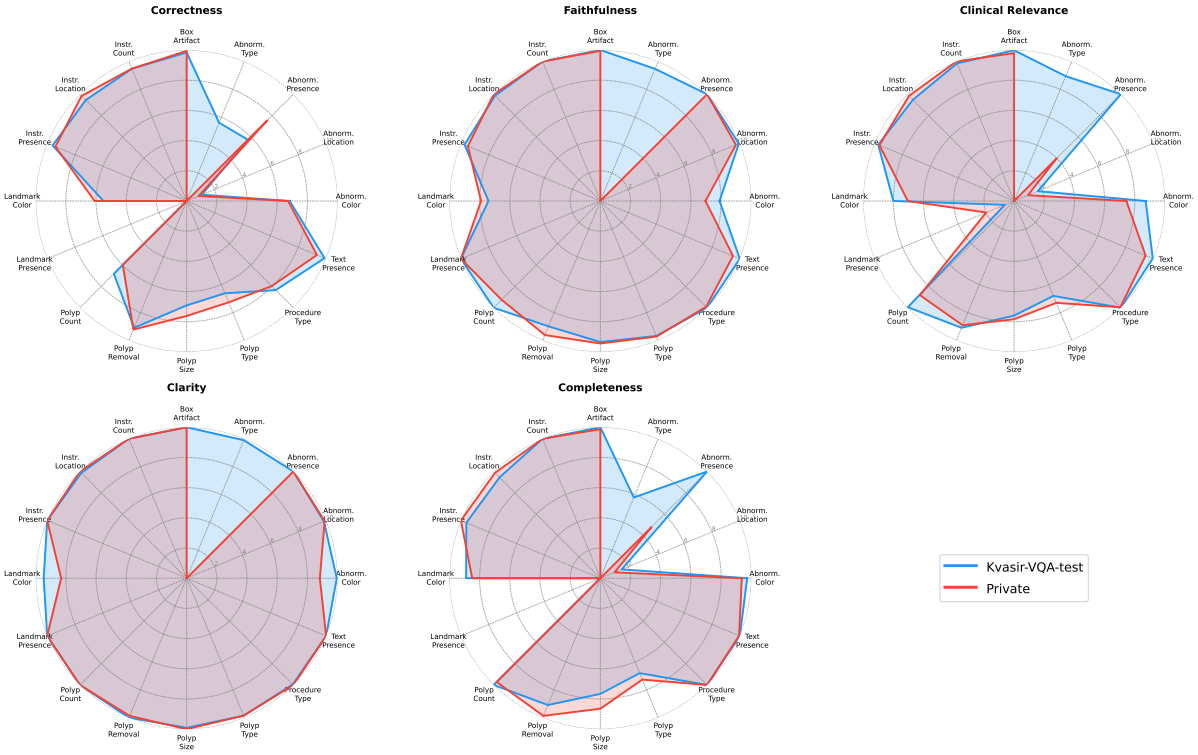
Table 7
Representative samples of successful and failed canonical answers produced by the Version 7 post-processor execution loop.

Input Question (Q)	Ground Truth (GT)	V7 Output (Pred)	Outcome & Root Cause Status
"How many instruments can be seen in the image?"	`1`	`1`	**Success:** Correct digit mapping via extraction regular expressions.
"What is the size of the polyp?"	`5-10mm`	`5-10mm`	**Success:** String normalization maps variations into canonical formats.
"Is there remaining polyp tissue after the procedure?"	`No`	`colonoscopy`	**Failure:** Branch-ordering defect; question mis-routed into procedure type loop.
"Where is the ulcerative colitis located?"	`Center-Right`	`none`	**Failure:** Cascade effect from active polyp-gate over-suppression rules.

The empirical validation of these structural behaviors is graphically mapped in the multi-dimensional radar visualizations shown in Figure 1.



(a) Mean LLM score by class



(b) Breakdown per LLM dimension

Figure 1: Radar plots comparing UIT_NEWRON performance on Kvasir-VQA-test vs. Private Test across clinical question classes and evaluation dimensions.

The following three failure modes were identified but not corrected within the competition timeline:

- **Multi-Panel Confusion.** Images containing split views occasionally caused the VLM to misallocate spatial attention, producing incorrect quadrant labels or conflated findings from different sub-panels.
- **Polyp–Anatomy Ambiguity.** Prominent mucosal folds and specular reflections were occasionally misclassified as active polyps, propagating errors into size, type, and removal queries.
- **Long-Tail Classification Sparsity.** Rare Paris classification sub-stages suffered from data scarcity, reducing precision relative to broader diagnostic categories.

7. Conclusion

We presented the UIT_NEWRON system for ImageCLEFmed-MEDVQA-GI-2026 Task 1, combining a QLoRA-fine-tuned Qwen2.5-VL-7B-Instruct backbone with a deterministic canonical post-processing pipeline. On the public validation set, the system achieved ROUGE-L of 0.8003 and METEOR of 0.4354, confirming that parameter-efficient quantisation combined with domain-specific output normalisation constitutes a practical baseline for clinical VQA on GI endoscopy data.

The primary contribution of this work, beyond the submitted system, is a detailed empirical characterisation of two failure modes broadly relevant to rule-based post-processing pipelines: (1) *branch-ordering sensitivity*, where evaluation order creates unintended routing side effects that are invisible on in-distribution validation data but surface dramatically under distribution shift; and (2) *context-gate fragility*, where a hard binary gating mechanism introduces a single point of failure that can cascade across a large fraction of outputs.

The severe gap between validation and test performance underscores that canonical post-processors must treat robustness under distribution shift as a first-class design requirement rather than an afterthought. Version 8 addresses both critical defects and provides a corrected baseline for future comparison.

This work extends our team's research program on efficient deep learning for medical imaging [1, 2, 9, 10, 19], now applied to the multimodal VQA setting for GI endoscopy.

Limitations The current system has three notable limitations. First, the rule-based post-processor is brittle under distribution shift, as demonstrated by the validation-to-test performance gap. Second, per-category evaluation relies on qualitative estimation due to incomplete logging during the competition window. Third, the QLoRA fine-tuning was conducted on a single question-type distribution and may not generalise to out-of-domain GI datasets without re-training.

Future work will pursue three directions: (1) replacing hard context gates with confidence-weighted soft routing using calibrated model probability estimates; (2) incorporating multiple independent VLM queries for critical binary classification flags (polyp presence, instrument presence) to reduce single-point failure risk through majority voting, building on our prior experience with genetic algorithm-enhanced ensemble methods for medical image classification [19]; and (3) integrating structured medical ontologies — including the Paris classification hierarchy [17] and SNOMED-CT anatomical terms [20] — into the post-processing layer to improve rare-class precision without increasing the risk of over-aggressive normalisation.

Code Availability The implementation used for the official competition submission is publicly available on Hugging Face.[1]

[1] https://huggingface.co/nhattan9999t/qwen2.5-vl-kvasir-vqa-submit

Acknowledgments

We thank the organisers of ImageCLEFmed-MEDVQA-GI-2026 for providing the evaluation framework and Kvasir-VQA dataset resources.

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

[1] T. B. Nguyen-Tat, T.-Q. T. Nguyen, H.-N. Nguyen, V. M. Ngo, Enhancing brain tumor segmentation in MRI images: A hybrid approach using UNet, attention mechanisms, and transformers, Egyptian Informatics Journal 27 (2024) 100528. doi:10.1016/j.eij.2024.100528.

[2] T. B. Nguyen-Tat, Q. H. Tran, T. N. Pham, V. M. Ngo, Evaluating pre-processing and deep learning methods in medical imaging: Combined effectiveness across multiple modalities, Alexandria Engineering Journal 119 (2025) 558–586. doi:10.1016/j.aej.2025.01.090.

[3] S. A. Hicks, S. Gautam, P. Halvorsen, M. A. Riegler, V. Thambawita, Overview of imageclefmedical 2026 – medical visual question answering for gastrointestinal tract, in: CLEF2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Berlin, Germany, 2026.

[4] 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, 2024. URL: https://arxiv.org/abs/2409.01437. doi:10.1145/3689096.3689458. arXiv:2409.01437.

[5] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-VQA-x1: A multimodal dataset for medical reasoning and robust MedVQA in gastrointestinal endoscopy, 2025. URL: https://arxiv.org/abs/2506.09958. arXiv:2506.09958.

[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[7] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708. doi:10.1109/CVPR.2017.243.

[8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016, pp. 457–468. doi:10.18653/v1/D16-1044.

[9] T. B. Nguyen-Tat, H.-A. Vo, P.-S. Dang, QMaxViT-Unet+: A query-based MaxViT-Unet with edge enhancement for scribble-supervised segmentation of medical images, Computers in Biology and Medicine (2025). doi:10.1016/j.compbiomed.2025.109762.

[10] T. B. Nguyen-Tat, L. Q. Truong, K. Q. Truong, T. H. Ngo, DoubleBlock-ViT: A MaxViT-based enhancement and dual-path skip connections for brain tumor segmentation in MRI scans, Computer Methods and Programs in Biomedicine 274 (2026) 109165. doi:10.1016/j.cmpb.2025.109165.

[11] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: Proceedings of the International Conference on Machine Learning, 2023, pp. 19730–19742. URL: https://arxiv.org/abs/2301.12597.

[12] W. Dai, J. Li, D. Li, A. Meng, S. Hoi, S. Savarese, C. Xiong, InstructBLIP: Towards general-purpose vision-language models with instruction tuning, in: Advances in Neural Information Processing Systems, 2023. URL: https://arxiv.org/abs/2305.06500.

[13] H. Liu, C. Li, Q. Wu, Y. J. Lee, LLaVA: Large language and vision assistant, in: Advances in Neural Information Processing Systems, 2023. URL: https://arxiv.org/abs/2304.08485.

[14] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang,

Z. Yang, H. Xu, J. Lin, Qwen2.5-VL technical report, 2025. doi:10.48550/arXiv.2502.13923. arXiv:2502.13923.
[15] X. Liu, Y. Zhang, W. Chen, et al., Detecting and evaluating medical hallucinations in large vision language models, arXiv preprint arXiv:2406.10185 (2024). URL: https://arxiv.org/abs/2406.10185.
[16] A. Ben Abacha, D. Demner-Fushman, H. Müller, et al., Overview of the medical visual question answering task (VQA-Med) at ImageCLEF 2019, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, CLEF 2019 Working Notes, CEUR-WS, 2019. URL: https://ceur-ws.org/Vol-2425/paper_147.pdf.
[17] P. in the Paris Workshop, The paris endoscopic classification of superficial neoplastic lesions: Esophagus, stomach, and colon, Gastrointestinal Endoscopy 58 (2003) S3–S43. doi:10.1016/S0016-5107(03)02159-X.
[18] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, in: Advances in Neural Information Processing Systems, 2023.
[19] T. B. Nguyen-Tat, T.-P. Nguyen-Duong, Optimizing knee osteoarthritis severity diagnostics: A GA-enhanced deep ensemble approach in medical imaging, Ain Shams Engineering Journal 16 (2025) 103524. doi:10.1016/j.asej.2025.103524.
[20] I. H. T. S. D. Organisation, Systematized nomenclature of medicine–clinical terminology (SNOMED CT), Journal of Medical Internet Research 25 (2023) e43750. URL: https://www.nlm.nih.gov/healthit/snomedct/. doi:10.2196/43750.

Generative AI Statement

During the preparation of this manuscript, generative AI tools were used exclusively for grammatical editing, structural formatting assistance, and LaTeX syntax checking. The experimental pipeline design, algorithm development, quantitative evaluation, and analysis of results were performed entirely by the human authors, who bear full scientific responsibility for the content and conclusions of this paper.