

Parameter-Efficient Vision–Language Fine-Tuning for Gastrointestinal Endoscopy VQA with Structured Explainability and Safety-Oriented Reasoning

Notebook for the CS_Morgan Lab at ImageCLEF2026

Raisa Nusrat Chowdhury^{1†}, Okib Ul Islam^{1†}, Ejiga Peter Ojonugwa Oluwafemi^{1,†}, Mahmudul Hoque^{1,†}, Ejiga Frederick Akor^{2†} and Md Mahmudur Rahman^{1,†}

¹Department of Computer Science, SCMNS, Morgan State University, Baltimore, Maryland, USA

²International Organization for Migration (IOM), Geneva, Switzerland

Abstract

Automated gastrointestinal (GI) endoscopy analysis has the potential to improve early disease detection and support more consistent clinical reporting; however, practical clinical deployment requires systems that are not only accurate but also interpretable and transparent about their uncertainty. We present the CS_Morgan submission to the ImageCLEFmed-MEDVQA-GI 2026 challenge, addressing clinically relevant visual question answering (Subtask 1) and multimodal explainability with safety-aware reasoning (Subtask 2). For Subtask 1, we develop a GI-VQA framework based on *Qwen2.5-VL-7B-Instruct*, adapted using 4-bit QLoRA and further fine-tuned from the publicly available SimulaMet Kvasir-VQA-x1 adapter, followed by a deterministic answer-normalization stage. For Subtask 2, we fine-tune *MedGemma-4B-IT* with QLoRA to generate a structured four-field response (ANSWER, JUSTIFICATION, CONFIDENCE, SAFETY_NOTE), using the Subtask 1 prediction as an auxiliary verification signal. To enhance interpretability, we integrate an external SAM segmenter fine-tuned on Kvasir-SEG to produce optional polyp masks, bounding boxes, and visual overlays as supporting evidence, while suppressing localization outputs for negative findings. Under the challenge’s LLM-as-a-Judge evaluation, the proposed system achieves strong Faithfulness (8.9/9.2 on the public/private sets) and Clarity (9.8/7.6), while Correctness and Completeness remain competitive on single-aspect questions but decline on more challenging multi-aspect (L2–L3) queries. We observe that the reported confidence and safety-note fields are now functioning as structured placeholders rather than calibrated or clinically verified signals. Notably, the model generally remains grounded in the available evidence rather than hallucinating unsupported information. We also report lexical scores (public ROUGE-1 = 0.52, ROUGE-L = 0.51), although, following the organizers’ recommendations, these metrics are considered secondary due to the variability of natural-language references. Finally, we discuss the current limitations of confidence estimation and safety supervision and outline future directions toward more reliable calibration and clinically trustworthy reasoning.

Keywords

Medical VQA, ImageCLEFmed 2026, Gastrointestinal Endoscopy, Vision-Language Models, QLoRA, Parameter-Efficient Fine-Tuning, Kvasir-VQA-x1, Explainability, Safety-Oriented Reasoning, MedGemma

1. Introduction

Colorectal cancer (CRC) remains one of the most frequently diagnosed cancers worldwide, accounting for approximately 1.9 million new cases each year [1]. Colonoscopy is widely regarded as the gold-standard screening procedure for early detection and diagnosis; however, polyp miss rates can reach up to 25% for small lesions, and adenoma detection rates (ADR) often vary substantially among endoscopists [2]. These challenges have driven growing interest in artificial intelligence (AI)-assisted systems that can improve diagnostic consistency, reduce inter-observer variability, support clinical reporting, and

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

[†] All authors contributed equally to this work.

✉ racho1@morgan.edu (R. N. Chowdhury); okis11@morgan.edu (O. U. Islam); ejiga.ojonugwa.peter@gmail.com (E. P. O. Oluwafemi); mahoq1@morgan.edu (M. Hoque); ejiga.fredrick@yahoo.com (E. F. Akor); md.rahman@morgan.edu (M. M. Rahman)

ORCID: 0000-0002-1663-6391 (R. N. Chowdhury); 0009-0000-8980-8273 (O. U. Islam); 0009-0003-2039-3075 (E. P. O. Oluwafemi); 0009-0006-5532-4135 (M. Hoque); 0009-0003-0700-6712 (E. F. Akor); 0000-0003-0405-9088 (M. M. Rahman)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

assist endoscopic decision-making.

Recent progress in medical vision-language models (VLMs) has enabled multimodal systems capable of jointly understanding medical images and natural-language queries [3, 4]. Within this context, medical visual question answering (Med-VQA) has emerged as a versatile framework that allows models to answer free-text clinical questions about an image, supporting tasks such as abnormality assessment, procedural understanding, anatomical landmark recognition, and instrument identification. The ImageCLEFmed-MEDVQA-GI 2026 challenge [5, 6] represents the fourth edition of the MEDVQA-GI benchmark and places particular emphasis on trustworthy AI behavior. Hallucinations, overconfident predictions, and misleading explanations are treated as key evaluation criteria through a dedicated behavioral safety framework. The challenge is based on the Kvasir-VQA-x1 dataset [7], which contains 159,549 GI endoscopy question-answer pairs spanning three reasoning complexity levels (L1/L2/L3). Two complementary subtasks are defined: Subtask 1 focuses on clinically relevant free-text answer generation, whereas Subtask 2 emphasizes multimodal explainability, confidence estimation, visual grounding, and safety-aware reasoning.

Our group has previously participated in MEDVQA-GI and related GI imaging challenges, investigating parameter-efficient VLMs for endoscopic VQA and medical image generation [8, 9, 10], as well as synthetic-data-driven approaches for classification and polyp segmentation [11, 12, 13]. Building on this experience, we present a practical and reproducible two-model framework rather than a novel algorithmic contribution:

- For Subtask 1, we continue QLoRA-based fine-tuning from the released SimulaMet Kvasir-VQA-x1 adapter on top of Qwen2.5-VL-7B-Instruct [14], followed by a deterministic normalization and answer-mapping stage that aligns model outputs with the dataset’s concise answer format.
- For Subtask 2, we fine-tune MedGemma-4B-IT [15] to generate a structured four-field output conditioned on the Subtask 1 prediction as an auxiliary verification signal, while incorporating an external Kvasir-SEG SAM segmenter [16, 17] for optional polyp localization with explicit negative-finding gating.
- We provide a transparent assessment of the proposed framework, including an analysis of the BLEU-4 artifact and a discussion of the current limitations of confidence estimation and safety supervision.

2. Related Work

Medical VQA: Medical visual question answering (Med-VQA) has evolved rapidly over the past decade, beginning with benchmark datasets such as VQA-RAD [18] and PathVQA [19]. In gastrointestinal (GI) endoscopy, the ImageCLEF MEDVQA-GI series has emerged as an important standard for assessing multimodal clinical reasoning. Kvasir-VQA [20] and its large-scale expansion, Kvasir-VQA-x1 [7], are extensively used resources that facilitate question answering across various levels of reasoning complexity. Early approaches relied on convolutional neural networks (CNNs) combined with attention-based multimodal fusion, whereas transformer-based architectures such as BERT [21] and Vision Transformers (ViT) [22] substantially improved cross-modal representation learning. Large vision-language models (VLMs), such as LLaVA-Med, InstructBLIP [23], Florence-2, and Qwen2-VL-based systems [24], have shown excellent skills in multimodal medical reasoning and answering open-ended clinical questions.

Remaining challenges: Despite these advances, GI endoscopy VQA remains a difficult problem due to the complexity of endoscopic imagery. Variations in mucosal appearance, lighting variations, specular reflections, motion blur, instrument occlusions, and visually identical pathology findings all add to the difficulty of reasoning. Existing VLMs struggle with multi-step procedural reasoning and multi-attribute clinical inference [10], and models trained on restricted clinical distributions usually display lower performance when applied to unexplored domains [25]. These limitations highlight the need for adaptation strategies that improve both robustness and generalization. We reexamine some of these issues, including multi-step and multi-attribute reasoning, in the context of our own pipeline

Parameter-efficient fine-tuning and adapter reuse: As the size of VLMs increases, parameter-efficient adaptation approaches become more crucial [26]. LoRA introduces trainable low-rank matrices into otherwise frozen transformer architectures [27], while QLoRA extends this approach through 4-bit quantization, enabling memory-efficient fine-tuning of billion-parameter models [28]. These strategies have shown high efficacy in several medical vision-language tasks, such as radiology report creation and picture captioning [29, 30]. A further benefit of current benchmark projects is the public distribution of task-specific adapters. In the case of Kvasir-VQA-x1, the availability of strong pretrained adapters allows downstream researchers to continue training from an already adapted checkpoint rather than restarting from a generic instruction-tuned model. This practical strategy significantly reduces computational costs and serves as the foundation of our approach in both challenge subtasks.

Interpretability, calibration, and safety: As medical AI systems get closer to clinical application, interpretability and dependability become more critical. Visualization methods like Grad-CAM [31] and structured reasoning approaches like chain-of-thought prompting [32] attempt to improve model predictions’ transparency and understandability. Confidence calibration is commonly evaluated using methods such as temperature scaling and Expected Calibration Error (ECE) [33], while segmentation frameworks including the Segment Anything Model (SAM) [16] and Kvasir-SEG [17] provide visual grounding that can improve the interpretability of model outputs. Retrieval-based techniques, generally combining sparse methods such as BM25 [34] with dense retrieval mechanisms, have been investigated to increase factual grounding, albeit retrieval is not used in the current study. The ImageCLEFmed-MEDVQA-GI 2026 challenge, which reflects the larger trend toward trustworthy medical AI, expressly examines hallucination, overconfidence, and clinically hazardous reasoning as fundamental system performance.

3. Task Overview and Dataset

The ImageCLEFmed-MEDVQA-GI 2026 challenge [5] targets VQA for GI endoscopy with emphasis on explainability, reliability, and safety-oriented reasoning, and explicitly addresses hallucinated outputs, misleading explanations, and overconfident predictions. **Subtask 1** (Clinically Relevant VQA) requires free-text answers for questions on abnormality detection, polyp characterization, instrument recognition, procedural understanding, and anatomical landmarks, evaluated with natural language generation (NLG) metrics including BLEU, ROUGE-L, METEOR, BERTScore-F1, and Token-F1. **Subtask 2** (Multimodal Explainability and Safety-Aware Reasoning) extends conventional VQA by requiring structured explanations, a reported confidence value, and clinically safe reasoning supported by visual evidence; organizer-side criteria additionally consider hallucination rate, Expected Calibration Error (ECE), overconfidence, interpretability, and clinical plausibility.

Both subtasks are based on Kvasir-VQA-x1 [7], a large-scale GI endoscopy Med-VQA benchmark with 159,549 clinically validated question-answer pairs divided into three complexity levels, from simple descriptive questions to multi-attribute compositional reasoning. The dataset covers abnormality detection, polyp characterization, procedural interpretation, anatomical landmark recognition, instrument localization, and artifact identification. The QA pairings are constructed using an LLM-assisted pipeline and expert validation. The reference responses are often brief, such as a single token or phrase, which affects how NLG metrics behave (Section 6.2).

4. Methodology

We use a two-stage pipeline (Figure 2). Subtask 1 produces a concise clinical answer; Subtask 2 consumes that answer, verifies and explains it, reports a confidence value and safety note, and optionally attaches visual evidence. Both stages adapt their backbone with 4-bit QLoRA initialized from the corresponding publicly released SimulaMet adapter, which we treat as a strong starting point rather than training from the base instruction model. Figure 1 gives a graphical overview of the end-to-end workflow, from endoscopic image acquisition to a clinician-facing report; Figures 2 and 3 then detail the two stages.

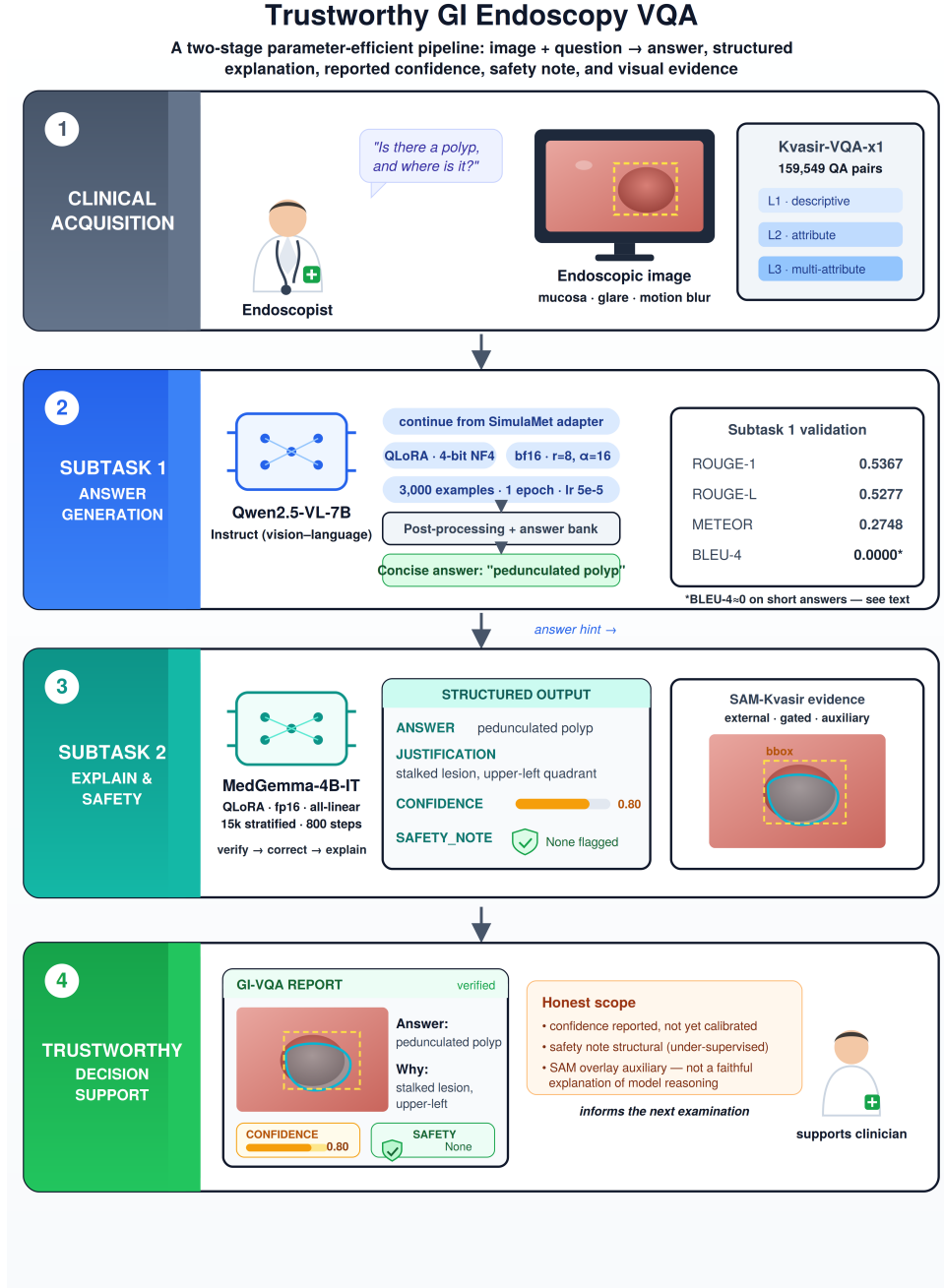


Figure 1: Graphical overview of the CS_Morgan pipeline. An endoscopic image and a clinical question enter a QLoRA-adapted Qwen2.5-VL-7B answer generator (Subtask 1); the answer is passed to a QLoRA-adapted MedGemma-4B model that produces a structured ANSWER/JUSTIFICATION/CONFIDENCE/SAFETY_NOTE response with gated SAM-Kvasir visual evidence (Subtask 2), yielding a clinician-facing report. The “honest scope” panel summarizes the current limitations of the confidence, safety, and grounding signals.

4.1. Clinically Relevant VQA (Subtask 1)

Backbone and adaptation. Subtask 1 model is built on Qwen2.5-VL-7B-Instruct [14], loaded using 4-bit NF4 quantization with double quantization and bfloat16 computation. Instead of initializing new LoRA adapters, we continue training from the publicly released SimulaMet Qwen Kvasir-VQA-x1 adapter, treating the released adapter weights as trainable parameters during QLoRA fine-tuning [27, 28]. The adapter configuration is retained from the original release, with rank $r=8$, scaling factor $\alpha=16$, and identical target modules. Each training example is represented as a single-image conversational turn of the form [Image, Question] → Answer.

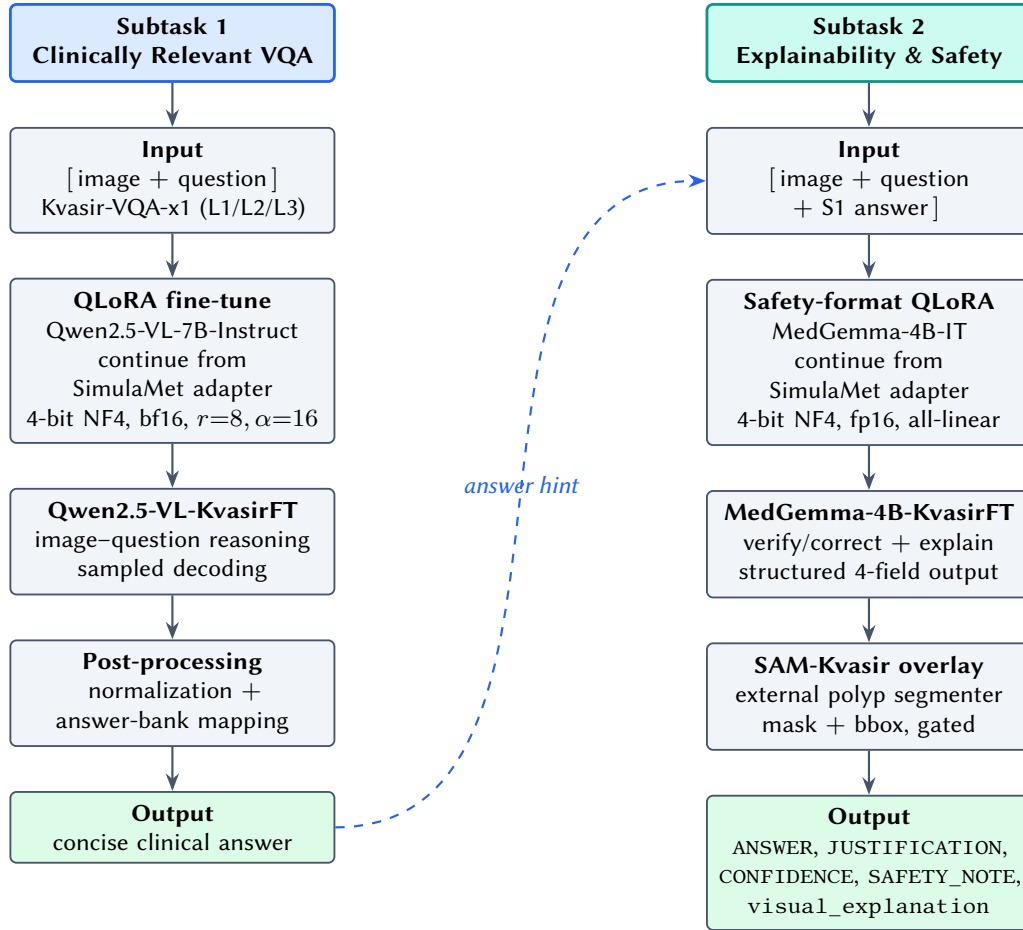


Figure 2: Two-stage CS_Morgan pipeline. Subtask 1 generates a concise answer with a QLoRA-adapted Qwen2.5-VL-7B model and a deterministic post-processing stage. The answer is passed as a verification hint to Subtask 2, where a QLoRA-adapted MedGemma-4B model produces a structured four-field response and an external Kvasir-SEG SAM segmenter optionally attaches visual evidence. Both adapters are initialized from the corresponding public SimulaMet checkpoints.

Training budget. To maintain a modest computational footprint, we perform one additional epoch of training on a randomly sampled subset of 3,000 Kvasir-VQA-x1 examples (seed 42). Training uses a per-device batch size of 1 with gradient accumulation of 8 (effective batch size 8), a learning rate of 5×10^{-5} , cosine learning-rate scheduling with 3% warmup, gradient checkpointing, and bf16 mixed precision. This setup is intended as lightweight continued adaptation of an already strong task-specific model rather than full retraining on the complete dataset.

Decoding. During inference, we follow the decoding settings recommended by the dataset authors : `max_tokens=512`, `temperature=0.3`, `top_k=20`, `top_p=0.7`, and `repetition_penalty=1.05`. We found that these settings produce more reliable responses than purely greedy decoding, which occasionally truncates or distorts short clinical answers.

Post-processing. Model outputs are passed through a deterministic normalization and answer-mapping pipeline that removes formatting artifacts, standardizes common expressions, and maps predictions to the concise answer style used throughout the dataset via a curated answer bank. Because the reference answers are typically short and highly standardized, this step provides a noticeable improvement in overlap-based evaluation metrics. The resulting system is referred to as *Qwen2.5-VL-KvasirFT*.

Prompting. For inference, we employ a strict label-only system prompt that encourages the model to generate the shortest medically accurate response, together with question-type-specific templates for yes/no, colour, count, and procedural queries. This strategy keeps predictions concise and helps

reduce hallucinated content, although it can introduce lexical mismatches when compared with the longer natural-language references used in the private test set (Section 6.2).

4.2. Multimodal Explainability and Safety-Aware Reasoning (Subtask 2)

Structured output contract. We fine-tune MedGemma-4B-IT [15] to generate a fixed four-field output:

- **ANSWER**: a concise clinically relevant response;
- **JUSTIFICATION**: 2–3 sentences grounded in visible evidence;
- **CONFIDENCE**: a float in $[0, 0.85]$ (capped during decoding);
- **SAFETY_NOTE**: a potential clinical concern, or None

Backbone and adaptation. We use MedGemma-4B-IT [15] loaded in 4-bit NF4 with a float16 compute dtype. Because MedGemma is trained in fp16, we retain fp16 rather than bf16 computation. Instead of initializing new adapters, we continue QLoRA training from the publicly released SimulaMet MedGemma Kvasir-VQA-x1 adapter using rank $r=8$, $\alpha=16$, dropout 0.05, and a 1-linear target modules. Training runs for 800 optimizer steps (capped within two epochs) on a complexity-balanced subset of 15,000 examples (5,000 from each of L1/L2/L3), with batch size 1, gradient accumulation 8, learning rate 5×10^{-5} , cosine scheduling, and gradient checkpointing. The resulting model is denoted *MedGemma-4B-KvasirFT*.

Verification-style conditioning. During inference, MedGemma is conditioned on the Subtask 1 answer as an auxiliary verification signal. When the Subtask 1 prediction is informative, the model is instructed to verify it against the image and revise it if necessary before generating an explanation; when the answer is empty or uninformative (e.g., “unknown”), the model responds directly. A lightweight selection rule retains the Subtask 1 answer when it is meaningful and otherwise falls back to MedGemma’s prediction. Outputs are parsed field-by-field using a tolerant regular-expression parser, while confidence values are clamped to $[0, 0.85]$ and default to 0.5 when parsing fails (Figure 3).

Justification robustness. Because generic justifications are common, we detect near-empty or boilerplate explanations and replace them with a question-aware template describing what evidence was examined while acknowledging when localized evidence is unavailable. Importantly, this template does not introduce new image findings; it serves as a transparency fallback rather than a source of clinical claims.

Visual evidence via an external segmenter. For optional visual grounding, we use an off-the-shelf SAM model [16] fine-tuned on Kvasir-SEG [17]. A trigger/negative-phrase rule determines whether visual evidence should be generated: outputs are produced only when the answer or justification mentions plausible visible targets such as polyps, lesions, instruments, landmarks, or artefacts, and are suppressed for explicit negatives (“no polyp”, “not visible”). When triggered, the module generates a segmentation mask, an overlay, and a bounding box derived from the mask’s largest contour. We emphasize that this segmenter operates independently of the VLM’s reasoning and therefore provides auxiliary visual context rather than a faithful explanation of the model’s internal attention.

5. Experimental Setup

Both subtasks were developed in Python with `ms-swift` 3.8.0, `transformers`, `peft`, `bitsandbytes`, and `qwen_vl_utils`, running on a single A100-class GPU with checkpoint caching and automatic resume. Table 1 summarizes the two training configurations as implemented. Subtask 1 submissions are produced and validated with the `medvqa` CLI against the `gi-2026` competition; Subtask 2 is packaged as a JSONL submission with optional visual assets in a ZIP archive.

6. Results and Evaluation

Following the challenge’s protocol, Subtask 1 is scored based on two sets: a public re-computation set called SimulaMet Kvasir-VQA-test, with 1,384 single-aspect complexity-1 examples, and the official

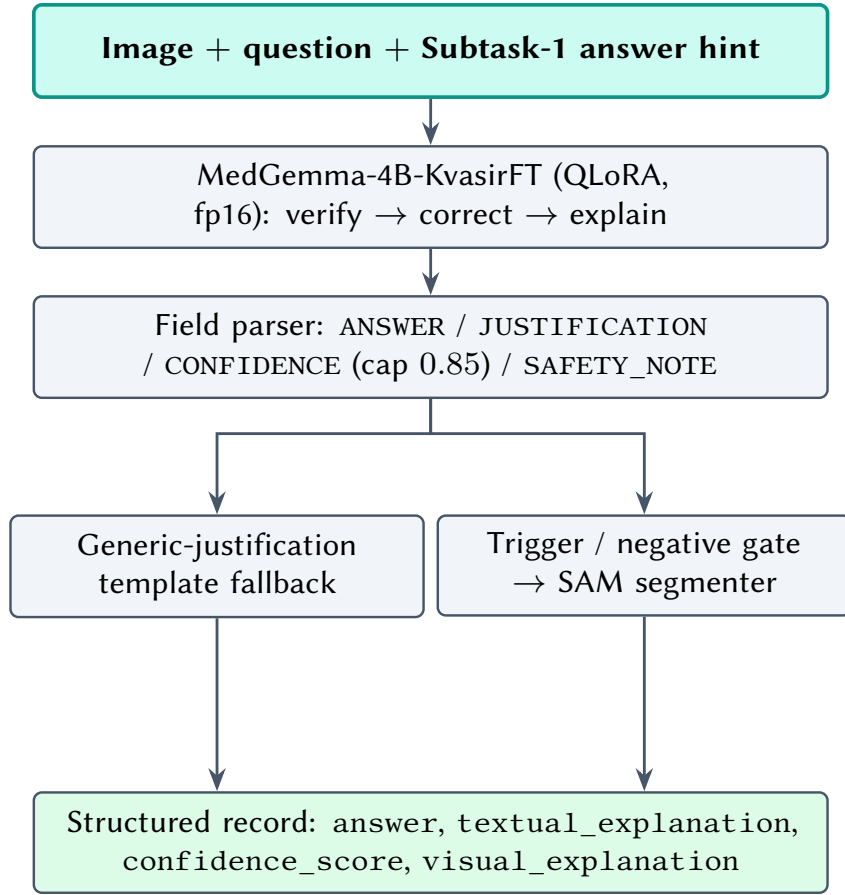


Figure 3: Subtask 2 inference flow. MedGemma produces a structured response that is parsed into fields; a template fallback repairs boilerplate justifications and a gated external SAM segmenter optionally adds visual evidence.

Table 1

Training configuration per subtask, as implemented.

	Subtask 1	Subtask 2
Backbone	Qwen2.5-VL-7B	MedGemma-4B-IT
Init adapter	SimulaMet Qwen FT	SimulaMet MedGemma FT
Quantization	4-bit NF4	4-bit NF4
Compute dtype	bf16	fp16
LoRA rank/ α	8 / 16 (inherited)	8 / 16
LoRA targets	inherited	all-linear
LoRA dropout	–	0.05
Train data	3k random	15k stratified (5k/level)
Epochs / steps	1 epoch	800 steps (≤ 2 ep)
LR / schedule	$5e-5$ / cosine	$5e-5$ / cosine
Eff. batch	8	8

private test set, Kvasir-VQA-x1-private, which includes 3,768 examples across complexity levels L1 to L3. Evaluations use two methods: surface-level lexical metrics and an LLM-as-a-judge. This judge uses Qwen3.6-27B to rate predictions on Correctness, Faithfulness, Clinical Relevance, Clarity, and Completeness, scoring them from 0 to 10. Since organizers say that lexical metrics don’t work well with the natural-language private references, we focus on the LLM-judge evaluation as our main results. We then discuss the lexical-metric caveat in Section 6.2.

6.1. LLM-judge evaluation

Table 2 shows the mean LLM-judge scores for both sets. The system is super faithful with scores of 8.88 (public) and 9.15 (private), meaning it barely adds unsupported info and the answers are clear, at 9.77 and 7.59. Correctness, Clinical Relevance, and Completeness are strong on the public set (7.19, 8.16, 8.11) but lower on the private one (5.35, 5.03, 5.06). As the complexity breakdown (Table 3) shows, this private drop is concentrated in the harder multi-aspect questions.

Table 2

Mean LLM-judge scores (0–10, higher is better) on the public (Kvasir-VQA-test) and private (Kvasir-VQA-x1) Subtask 1 sets.

Dimension	Public	Private
Correctness	7.19	5.35
Faithfulness	8.88	9.15
Clinical Relevance	8.16	5.03
Clarity	9.77	7.59
Completeness	8.11	5.06

On the private set, Faithfulness remains high (9.69 → 9.12 → 8.57 for L1→L2→L3), but Correctness (8.43 → 4.95 → 2.31), Clinical Relevance and Completeness (9.15 → 4.18 → 1.39) drop sharply with increasing complexity. This shows the model stays grounded and doesn’t hallucinate info. Yet, increasingly *omits* required components of multi-part questions rather than fabricating them.

Table 3

LLM-judge scores by question complexity. The public set contains only complexity-1 (single-aspect) questions; the private set spans L1 (single-), L2 (dual-), and L3 (triple-aspect) questions.

Dimension	Public	Private		
	L1	L1	L2	L3
Correctness	7.19	8.43	4.95	2.31
Faithfulness	8.88	9.69	9.12	8.57
Clinical Relevance	8.16	8.63	4.20	1.87
Clarity	9.77	9.78	7.11	5.62
Completeness	8.11	9.15	4.18	1.39

Per question class. In Figures 4 and 5, the judge scores are broken down into the sixteen single-aspect question classes. The system performs best on presence, counting questions, and text or box artifact detection, like Instrument Presence, Count, Text Presence, Box Artifact, and Abnormality Presence rates high for both Faithfulness and Clarity. On the flip side, it struggles with fine-grained typing tasks, particularly with Abnormality Type. Plus, there are two odd cases due to annotation-style differences. For instance, Polyp Removal status scores way higher on the private set compared to the public one. Similarly, Landmark Presence shows low Correctness but tops Faithfulness; this means the model’s responses, while conservative, don’t quite match up with what the judges expect based on the reference provided.

6.2. Lexical metrics

We also include standard lexical metrics in Table 4, though they need careful interpretation. The organizers point out that surface-overlap scores aren’t reliable for this task, especially on the private set. References there use full sentences, whereas our system spits out concise label-style answers. For the public set, which has shorter references, the system gets ROUGE-1 equal to 0.5160 and ROUGE-L at 0.5052. But on the private set, the same answers fare much worse; ROUGE-1 drops to 0.1987 and ROUGE-L to 0.1916, mostly due to the label-vs.-sentence issue, not clinical correctness, as indicated by

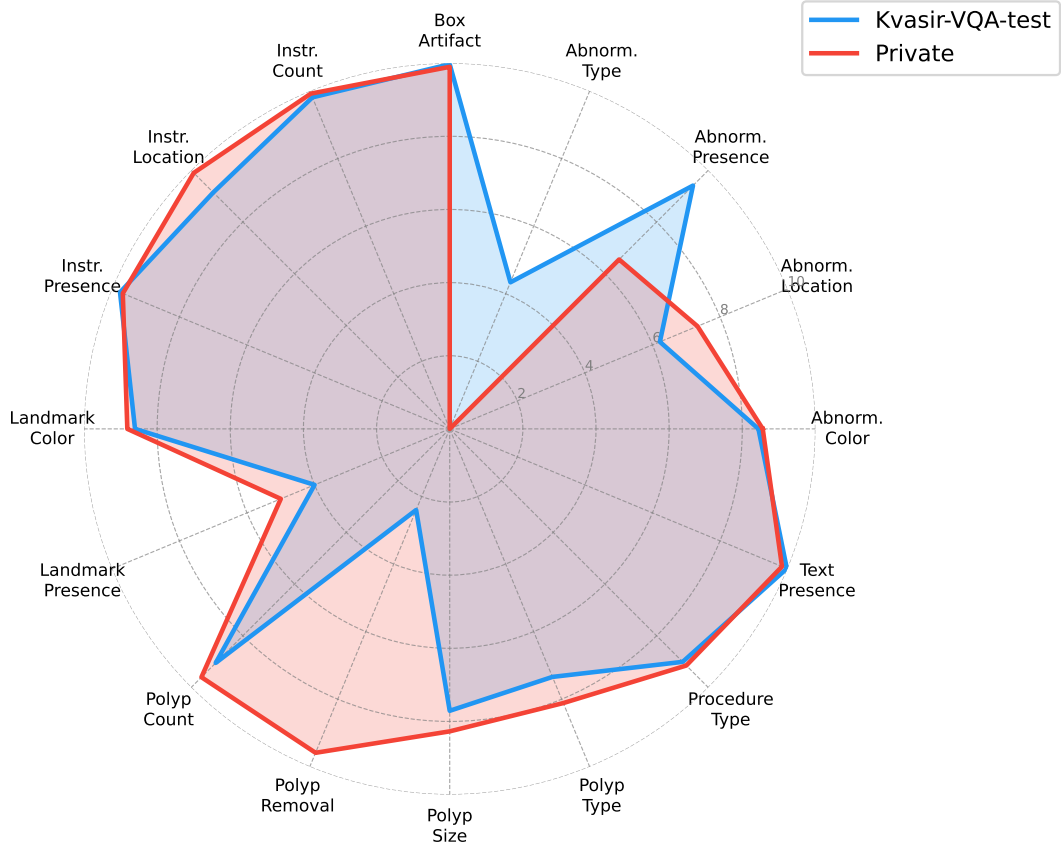


Figure 4: Mean LLM-judge score by question class (public vs. private).

the high Faithfulness in Table 2. This discrepancy hits BLEU scores harder since it zeroes in on n -grams. Many refs in this set are single words or brief phrases, meaning higher-order BLEU struggles to match up n -grams, causing BLEU-4 to plummet to almost zero due to its geometric mean formulation that tanks when p_2 , p_3 , and p_4 all hit 0. Even BLEU-1 lands at just 0.1918 for the public dataset. Given these issues, we consider the LLM-judge dimensions a better quality indicator for Subtask 1.

Table 4

Lexical metrics on the public (Kvasir-VQA-test) and private (Kvasir-VQA-x1) Subtask 1 sets. These surface-overlap scores understate clinically correct answers phrased differently from the reference and are reported for completeness.

Metric	Public	Private
BLEU-1	0.1918	0.0035
BLEU-2	0.1185	0.0022
ROUGE-1	0.5160	0.1987
ROUGE-2	0.0095	0.0232
ROUGE-L	0.5052	0.1916
METEOR	0.2594	0.0864

6.3. Subtask 2 Submission Integrity

We report the structural validation checks done by the build pipeline before packaging for Subtask 2, since the organizers score it based on the submitted JSONL file (Table 5). Each record is checked for the right fields, a unique and consecutive `val_id`, a `confidence_score` within the allowed range, a

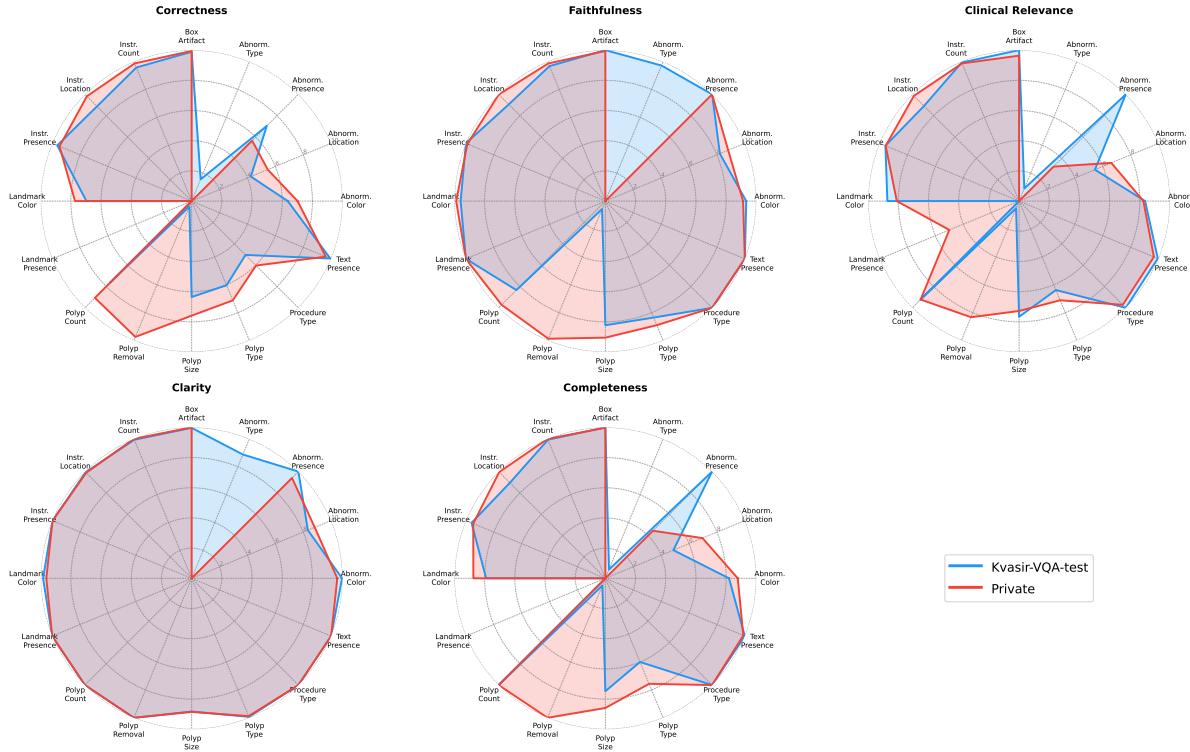


Figure 5: LLM-judge scores by question class, one panel per dimension (Correctness, Faithfulness, Clinical Relevance, Clarity, Completeness), comparing the public (blue) and private (red) sets.

list-typed `visual_explanation`, and paths to resolvable visual assets. If any check fails, the archive gets rejected. These checks ensure submission format integrity, not the quality metrics, which the organizers handle separately.

Table 5

Subtask 2 submission-format validation enforced by the build pipeline.

Validation check	Result
Required fields present in every record	100% (enforced)
<code>val_id</code> unique and contiguous ($0 \dots N-1$)	Yes (enforced)
<code>confidence_score</code> within $[0, 0.85]$	100% (clamped)
<code>visual_explanation</code> is a list in every record	100% (enforced)
Referenced visual files present in archive	All (0 missing)
Negative-finding gating applied to visuals	Yes

7. Ablation Studies

We conducted controlled experiments, focusing on just one variable at a time, to figure out how individual design choices impact the results. Each variant used the same setup for adapter and normalization processes along with the answer bank, excluding the element being tested. This way, any changes in scores were clearly linked to that particular factor.

Decoding configuration. Holding the model and post-processing fixed, we compared greedy decoding (temperature= 0, top_k= 1, short max_tokens) against the dataset authors’ recommended sampling configuration (temperature= 0.3, top_k= 20, top_p= 0.7, repetition_penalty= 1.05, max_tokens= 512). A length-distribution diagnostic confirmed that the sampling configuration did not induce prose drift: predicted answer lengths remained close to the short reference distribution.

Adapter initialization. Continuing with QLoRA from the SimulaMet adapter, instead of starting with new adapters on the base instruction model, is crucial for achieving reasonable overlap scores with a limited training budget. This is what makes a 3,000 example run workable.

Post-processing. Disabling the normalization and answer-bank stage lowers overlap metrics since the raw outputs differ from the dataset’s very standardized short-answer format, even if the responses are actually correct. Because full official validation wasn’t completed for every version, we note the observed effects of each design choice in Table 6. We’ll save thorough testing of each variant for future research.

Table 6

Qualitative ablation of the Subtask 1 design choices.

Variant	Expected effect	Observation
Greedy decoding	Shorter outputs	More truncation / distortion
Sampling decoding	Better answer completion	Short answer length preserved
No normalization	Lower lexical overlap	More canonical mismatch
SimulaMet initialization	Better small-budget adaptation	Enables viable 3k-example training

8. Discussion

The pipeline achieves moderate overlap scores on Subtask 1 and produces well-formed, parseable structured outputs for Subtask 2. However, several limitations warrant clear acknowledgment.

Confidence is currently reported but not calibrated In the present Subtask 2 supervision setting, the confidence target is fixed, leading the fine-tuned model to converge toward a near-constant output. To maintain validity, we clamp this value to $[0, 0.85]$ during decoding. While this ensures a syntactically valid confidence field, it does not reflect true probabilistic calibration. Notably, we do not yet apply post-hoc calibration techniques such as temperature scaling, nor do we report metrics like Expected Calibration Error (ECE) [33]. Achieving meaningful calibration will require supervision with variable, outcome-dependent targets, alongside an explicit calibration step. We identify this as a high-priority direction for future work.

Safety notes remain under-supervised. Similarly, the safety-note target is fixed (None) during training, which discourages the model from surfacing potential concerns at inference time. While the field establishes a structural placeholder for safety-oriented reasoning, its effective use requires training data enriched with meaningful safety annotations.

Justifications combine learned behavior and templating. Given that the training targets for justification are generic, the quality of generated explanations relies substantially on prompt design and template-based fallbacks, rather than consistently on learned, image-grounded reasoning. Although the fallback mechanism is intentionally transparent, it does not provide localized or evidence-specific explanations.

Visual grounding is auxiliary. The SAM-based overlay operates as an external segmentation module, independent of the VLM’s internal reasoning process. It serves as a practical sanity check and offers human-interpretable region proposals; however, it should not be interpreted as a faithful explanation of the model’s decision-making, nor do we claim it mitigates textual hallucination.

Answer style influences the public-private performance gap. The system adopts a strict label-only prompting strategy alongside question-type-aware templates, promoting concise and faithful outputs that align well with compact reference answers (as in the public set). However, on the private set, where references are more descriptive, this brevity reduces lexical overlap and negatively impacts judged Completeness, even when predictions are correct. Incorporating both concise and descriptive answer targets during training represents a direct and high-priority avenue for improvement.

Performance degrades on multi-aspect questions. Both judged Correctness and Completeness decline markedly as question complexity increases from single- to multi-aspect queries, while Faithfulness remains stable. This pattern suggests omission of required components rather than hallucination. Addressing this limitation will likely require explicit multi-attribute supervision and answer templates that systematically enumerate all requested aspects.

These limitations do not diminish the core engineering contribution—namely, an efficient, reproducible, and well-structured pipeline but they do circumscribe the strength of any trustworthiness claims. We therefore scope our conclusions accordingly.

9. Conclusion and Future Work

We presented the CS_Morgan submission to ImageCLEFmed-MEDVQA-GI 2026: a two-stage pipeline that pairs a QLoRA-adapted Qwen2.5-VL-7B answer generator with a QLoRA-adapted MedGemma-4B structured explainer, both continued from public SimulaMet adapters, plus an optional external Kvasir-SEG segmenter for visual evidence. Under the challenge’s LLM-judge evaluation the system is highly faithful (8.88/9.15 on the public/private sets) and clear, with answer quality concentrated on single-aspect questions and degrading on harder multi-aspect questions; following the organizers’ guidance we de-emphasize lexical metrics. For future improvements, we plan to work on varied, result-linked confidence feedback and better post-hoc calibration. We also want to add real safety annotations and richer visual justification. Additionally, we’re looking to assess the visual evidence part more closely, like comparing it with Kvasir-SEG using IoU and Dice scores. And lastly, we intend to practice with different kinds of answer targets, along with explicit multi-attribute guidance, to narrow down the gaps between the public and private sets and handle multi-aspect.

10. Declaration on Generative AI

During the preparation of this work, the authors utilized generative AI tools for proofreading, grammar correction, and enhancing the clarity and readability of the manuscript. All AI-generated suggestions were thoroughly reviewed and edited by the authors to ensure accuracy and integrity. The authors take full responsibility for the content of the final manuscript.

Acknowledgments

This work used computational resources provided by the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program under Allocation Request **CIS251242**, including **53.0 GPU hours on the NCSA Delta GPU system**. The authors acknowledge the National Science Foundation (NSF), ACCESS, and NCSA Delta for supporting the computational resources used in this research.

11. Online Resources

GitHub: <https://github.com/Ejigsonpeter/VQA-with-Structured-Explainability-and-Safety-Oriented-Reasoning>

References

- [1] I. Barua, D. G. Vinsard, H. C. Jodal, M. Løberg, M. Kalager, Ø. Holme, M. Misawa, M. Bretthauer, Y. Mori, Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis, *Endoscopy* 53 (2021) 277–284. doi:10.1055/a-1201-7165.

- [2] L. F. Sánchez-Peralta, L. Bote-Curiel, A. Picón, F. M. Sánchez-Margallo, J. B. Pagador, Deep learning to find colorectal polyps in colonoscopy: A systematic literature review, *Artificial Intelligence in Medicine* 108 (2020) 101923. doi:10.1016/j.artmed.2020.101923.
- [3] R. N. Chowdhury, M. Hoque, M. R. Hasan, E. P. O. Oluwafemi, M. M. Rahman, Modality-guided radiology caption prediction with small vision-language models and image classifier, 2025.
- [4] M. Hoque, M. Hasan, M. Emon, F. Khalifa, M. Rahman, Medical image interpretation with large multimodal models, in: *CEUR workshop proceedings*, CEUR Workshop Proceedings 3740, CEUR-WS.org 2024, 2024.
- [5] 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.
- [6] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ștefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryilmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.
- [7] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-vqa-x1: A multimodal dataset for medical reasoning and robust medvqa in gastrointestinal endoscopy, arXiv:2506.09958, 2025. arXiv:2506.09958.
- [8] O. O. Ejiga Peter, F. Akor Ejiga, F. Khalifa, M. M. Rahman, Parameter-efficient vlms for gastrointestinal endoscopy: Medical image generation and clinical visual question answering, in: *2025 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2025, pp. 1–7. doi:10.1109/bhi67747.2025.11269532.
- [9] O. O. Ejiga Peter, M. M. Rahman, F. Khalifa, Advancing ai-powered medical image synthesis: Insights from medvqa-gi challenge using clip, fine-tuned stable diffusion, and dream-booth + lora, arXiv:2502.20667, 2025. arXiv:2502.20667.
- [10] O. O. Ejiga Peter, M. Hoque, F. Akor Ejiga, R. N. Chowdhury, Solving medical data limitations through ai: Multi-modal vision-language learning for gastrointestinal vqa and synthetic training data generation, in: *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF)*, 2025.
- [11] O. O. Ejiga Peter, O. T. Adeniran, J.-O. A. MacGregor, F. Khalifa, M. M. Rahman, Text-guided synthesis in medical multimedia retrieval: A framework for enhanced colonoscopy image classification and segmentation, *Algorithms* 18 (2025) 155. doi:10.3390/a18030155.
- [12] O. O. Ejiga Peter, O. G. Akingbola, C. R. Amalahu, O. Adeniran, F. Khalifa, M. M. Rahman, Synthetic data-driven multi-architecture framework for automated polyp segmentation through integrated detection and mask generation, in: *Medical Imaging 2025: Clinical and Biomedical Imaging*, 2025, p. 78. doi:10.1117/12.3049369.
- [13] O. O. Ejiga Peter, B. D. Tunde, M. Hoque, D. Briggs, M. M. Rahman, F. Khalifa, Colonoscopy image synthesis: Transforming medical image classification with enhanced quality and computational efficiency, in: *2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI)*, 2025, pp. 1–6. doi:10.1109/icmi65310.2025.11141284.
- [14] Qwen Team, Qwen2.5-vl technical report, 2025. arXiv:2502.13923.
- [15] Google DeepMind Health AI Team, Medgemma technical report, 2025. arXiv:2507.05201.
- [16] 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. arXiv:2304.02643.
- [17] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen,

- Kvasir-seg: A segmented polyp dataset, in: MultiMedia Modeling (MMM), 2020, pp. 451–462. doi:10.1007/978-3-030-37734-2_37.
- [18] 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) 180251. doi:10.1038/sdata.2018.251.
 - [19] X. He, Y. Zhang, L. Mou, E. Xing, P. Xie, Pathvqa: 30000+ questions for medical visual question answering, 2020. *arXiv:2003.10286*.
 - [20] 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: *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications*, ACM, 2024, pp. 3–12. doi:10.1145/3689096.3689458.
 - [21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. *arXiv:1810.04805*.
 - [22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2020. *arXiv:2010.11929*.
 - [23] 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. *arXiv:2305.06500*.
 - [24] M. Hoque, M. R. Hasan, I. S. Emon, F. Khalifa, M. M. Rahman, Medical image interpretation with large multimodal models, 2024. Manuscript in preparation.
 - [25] S. Gaihre, A. T. Magar, P. Pokharel, L. Tiwari, Cs_morgan at imageclefmedical 2025: Gi endoscopy visual question answering, in: *CLEF 2025 Working Notes*, 2025.
 - [26] M. Hoque, F. Kim, Generative modeling of additive manufacturing lack-of-fusion defects using three-dimensional adversarial network (2026).
 - [27] 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. *arXiv:2106.09685*.
 - [28] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, 2023. *arXiv:2305.14314*.
 - [29] M. Hoque, R. N. Chowdhury, M. R. Hasan, O. O. Ejiga Peter, F. Khalifa, M. M. Rahman, An empirical evaluation of low-rank adapted vision-language models for radiology image captioning, *Bioengineering* 12 (2025) 1330. doi:10.3390/bioengineering12121330.
 - [30] R. N. Chowdhury, M. Hoque, M. R. Hasan, O. O. Ejiga Peter, M. M. Rahman, Modality-guided radiology caption prediction with small vision-language models and image classifier, 2025. Preprint.
 - [31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, *International Journal of Computer Vision* 128 (2020) 336–359. doi:10.1007/s11263-019-01228-7.
 - [32] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. *arXiv:2201.11903*.
 - [33] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, 2017. *arXiv:1706.04599*.
 - [34] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, *Foundations and Trends in Information Retrieval* 3 (2009) 333–389. doi:10.1561/15000000019.