From Perception to Explanation: A Multimodal Framework for GI Visual Question Answering

Krishna Tewari[1], Sukomal Pal[1]

[1]Indian Institute of Technology (BHU) Varanasi, India

Abstract

Gastrointestinal visual question answering has emerged as a promising approach for assisting clinicians in interpreting endoscopic images through multimodal reasoning. In this paper, we describe our participation in the ImageCLEFmed MEDVQA-GI 2026 challenge, addressing both clinically relevant VQA and multimodal explainability. For answer generation, we propose a lightweight multimodal framework that combines InceptionV3 for visual feature extraction, BioClinicalBERT for clinical question understanding, and BioGPT for autoregressive answer generation through multimodal prefix conditioning. To enhance transparency and reliability, we extend the framework with Grad-CAM-based visual grounding, token-level confidence estimation, Monte Carlo dropout uncertainty quantification, semantic alignment analysis, and BioGPT-based explanation generation. Experimental evaluation on the official benchmark demonstrates competitive performance, achieving a METEOR score of 0.4800 on the private test set and a completeness score of 8.22, while maintaining correctness and clinical relevance scores of 6.49 and 6.12, respectively. The results highlight the effectiveness of lightweight and explainable multimodal architectures for trustworthy GI endoscopy question answering and clinical decision support.

Keywords

Medical VQA, GI Endoscopy, Multimodal Explainability, Multimodal Medical AI

1. Introduction and Related Work

Gastrointestinal (GI) diseases represent a substantial global healthcare burden, encompassing conditions such as colorectal cancer, gastric cancer, inflammatory bowel disease, GI bleeding, and precancerous lesions. According to recent global cancer statistics, colorectal and gastric cancers remain among the leading causes of cancer-related mortality worldwide, highlighting the importance of early diagnosis and timely clinical intervention [1]. Endoscopic procedures, including colonoscopy, gastroscopy, and capsule endoscopy, serve as the primary diagnostic tools for detecting GI abnormalities. However, these procedures generate large volumes of high-resolution visual data that require expert interpretation, making diagnosis labor-intensive, time-consuming, and susceptible to inter-observer variability [2]. Consequently, the development of automated and clinically reliable computer-aided diagnostic systems has become an active area of research.

Recent advances in deep learning and multimodal artificial intelligence have significantly improved automated medical image analysis. In particular, Visual Question Answering (VQA) has emerged as a promising paradigm for interactive clinical decision support by combining image understanding with natural language reasoning [3, 4]. In a medical VQA setting, a model receives an image (I) and a clinical question (Q), and generates an answer (A) through the mapping $f : (I, Q) \rightarrow A$. Unlike conventional image classification tasks, VQA requires joint reasoning over visual and textual modalities, enabling clinicians to query specific findings, anatomical structures, procedural details, or pathological characteristics directly from medical images. The emergence of large vision-language models and transformer-based architectures has further accelerated progress in this area by enabling richer multimodal representations and improved cross-modal understanding [5, 6, 7, 8, 9].

The medical VQA field has evolved through several benchmark datasets and challenge initiatives. Early datasets such as VQA-RAD [3] and PathVQA [4] established foundational benchmarks for radiological and pathological reasoning, demonstrating the feasibility of multimodal clinical question answering.

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

krishnatewari.rs.cse24@iitbhu.ac.in (K. Tewari); spal.cse@iitbhu.ac.in (S. Pal)

0009-0005-6599-9956 (K. Tewari); 0000-0001-8743-9830 (S. Pal)

Within the GI domain, the release of HyperKvasir [2], Kvasir-VQA [10], and the recently introduced Kvasir-VQA-x1 dataset [11] has facilitated the development of specialized GI-VQA systems capable of reasoning about endoscopic findings and disease characteristics. Concurrently, the ImageCLEFmed-MEDVQA challenges have played a crucial role in advancing multimodal medical reasoning by providing standardized evaluation benchmarks and promoting reproducible research in clinical vision-language understanding [12, 13, 14, 15, 16]. The 2026 edition of the challenge further emphasizes clinically grounded reasoning, safety-aware predictions, confidence estimation, and explainability, reflecting the growing demand for trustworthy AI systems in healthcare [17, 18].

Parallel developments in biomedical language modeling have contributed significantly to multimodal medical AI. Domain-specific language models such as BioBERT [19], ClinicalBERT [20], and BioGPT [21] have demonstrated strong performance across a variety of biomedical natural language processing tasks by leveraging large-scale medical corpora. These models provide contextual representations that capture specialized clinical terminology and semantic relationships, making them highly suitable for medical VQA applications. Furthermore, parameter-efficient adaptation techniques such as LoRA [22] have enabled the practical deployment of large foundation models within resource-constrained environments while preserving competitive performance.

Despite substantial improvements in answer accuracy, recent research increasingly recognizes that predictive performance alone is insufficient for clinical deployment. Medical AI systems must also provide interpretable and trustworthy explanations that allow clinicians to understand, verify, and assess model predictions [23, 24, 25]. Visual explanation techniques such as Grad-CAM [26] have become widely adopted for highlighting image regions responsible for model decisions, while uncertainty estimation methods including Monte-Carlo Dropout [27] and calibration techniques [28] help quantify predictive confidence and mitigate overconfident errors. These capabilities are particularly important in high-stakes medical environments where erroneous predictions may directly influence patient care.

Recent studies have explored explainable multimodal frameworks that integrate visual grounding, segmentation-based localization, and confidence-aware reasoning within medical question-answering systems. In dermatological and GI imaging, CLIPSeg-based segmentation, Florence-2 vision-language models, and multimodal explanation pipelines have demonstrated promising results for generating clinically interpretable predictions [29, 30]. Furthermore, confidence-aware explanation generation has been shown to improve transparency by coupling model predictions with uncertainty estimates and textual rationales [31]. These developments indicate a broader transition from purely predictive architectures toward trustworthy multimodal systems capable of supporting clinical decision-making.

Motivated by these advances and the objectives of the ImageCLEFmed-MEDVQA-GI 2026 challenge, we propose a multimodal GI VQA framework that combines visual representations extracted from InceptionV3, clinical language understanding through BioClinicalBERT, and answer generation using BioGPT. To improve reliability and interpretability, the proposed system incorporates Grad-CAM-based visual grounding, Monte-Carlo dropout uncertainty estimation, token-level confidence analysis, and image-question semantic alignment measures. By integrating multimodal reasoning with confidence-aware explanations, our approach aims to provide clinically meaningful answers while supporting transparency and trustworthiness in AI-assisted GI diagnosis.

The rest of the paper is structured as follows: Section 2 describes the task overview and dataset employed; Section 3 presents the proposed methodology for each subtask; Section 4 reports results; and Section 5 concludes with key findings.

2. Task Overview and Dataset

The ImageCLEFmed MEDVQA-GI 2026 challenge [18] as part of ImageCLEF [17] benchmark, focuses on advancing trustworthy and clinically grounded VQA systems for GI endoscopy. The challenge aims to promote the development of multimodal vision-language models capable of understanding endoscopic images, answering clinically relevant questions, and providing transparent explanations to support their predictions.

Task 1: Clinically Relevant Visual Question Answering. The primary objective of this task is to generate accurate answers to questions associated with GI endoscopic images. Given an image-question pair, participating systems must infer clinically meaningful information from the visual content and produce concise answers. The questions cover a wide range of topics, including anatomical structures, pathological findings, procedural observations, and diagnostic interpretations.

Task 2: Multimodal Explainability and Safety-Aware Reasoning. This task extends conventional VQA by requiring models to provide explanations that justify their predictions. In addition to generating an answer, systems are expected to produce textual reasoning and visual evidence highlighting image regions that support the decision-making process. The task further evaluates the reliability of explanations and encourages the development of models that reduce hallucinations, improve interpretability, and provide clinically trustworthy outputs. This reflects real-world clinical requirements where transparency and accountability are essential for the adoption of AI-assisted diagnostic systems.

2.1. Dataset

The challenge dataset (Kvasir-VQA-x1 [11]) is based on an expanded version of the Kvasir-VQA [10] benchmark. It comprises 6,500 GI endoscopic images from HyperKvasir [2] and Kvasir-Instrument [32], paired with 159,549 QA pairs stratified by reasoning complexity. Each sample in the dataset consists of:

- A gastrointestinal endoscopic image,
- A clinically relevant natural-language question,
- A corresponding ground-truth answer.

A sample images from the dataset are shown in Figure 1. While the sample question-answer pairs along with varying complexity are shown in Table 1.

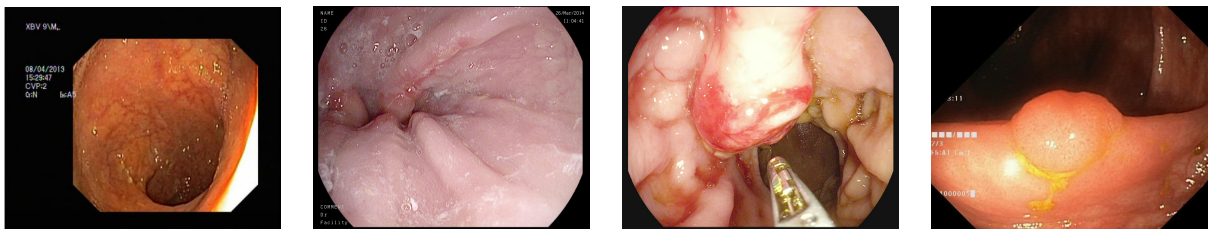

Figure 1: Sample images from the dataset. The four examples illustrate the diversity of visual appearances encountered in the collection, including variations in anatomical structures, illumination conditions, and viewing angles.

Table 1
Examples of questions with different complexity levels from the dataset.

Complexity	Question	Answer	Question Class
1	Which anatomical landmark is visible in the image?	No identifiable anatomical landmark present	landmark_location
2	What procedure is depicted in the image and what colors are associated with the abnormality?	Evidence of colonoscopy findings with pink and red mucosal lesions	procedure_type, abnormality_color
3	Are there any anatomical landmarks visible, what type of polyps are present, and what colors are the observed abnormalities?	No anatomical landmarks identified, no polyps observed, and multiple abnormalities with pink and red coloration.	landmark_presence, polyp_type, abnormality_color

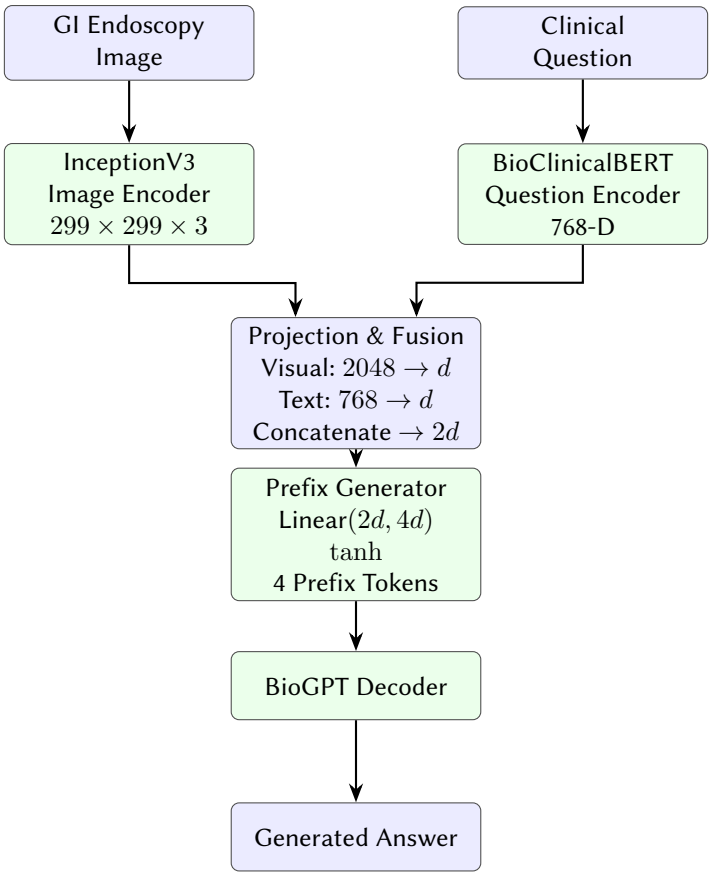

Figure 2: Task 1 framework for GI VQA. Visual features extracted by InceptionV3 and textual embeddings generated by BioClinicalBERT are concatenated, projected into learnable prefix embeddings, and used to condition the BioGPT for answer generation.

3. Methodology

This section describes our end-to-end methodology for both the subtasks of the ImageCLEFmedical MedVQA 2026 challenge.

3.1. Subtask 1

Figure 2 presents the overall architecture of the proposed GI-VQA framework. The framework consists of three major components: (i) a visual encoder for extracting endoscopic image representations, (ii) a medical language encoder for question understanding, and (iii) a generative language model that produces the final answer through multimodal prefix conditioning.

Images are resized to the input resolution required by the visual encoder and normalized using ImageNet preprocessing statistics. Clinical questions are tokenized using the BioClinicalBERT tokenizer, while answer sequences are tokenized using the BioGPT tokenizer. During training, answer tokens serve as supervision signals for autoregressive language modeling.

3.1.1. Visual Feature Extraction

The visual branch employs the InceptionV3 architecture pre-trained on ImageNet as the image encoder. The final classification layer is removed, and the output of the global pooling layer is used as a compact visual representation. Given an input endoscopic image I, the visual encoder extracts a 2048-dimensional feature vector: $\mathbf{v} = f_{img}(I)$, where $f_{img}(\cdot)$ denotes the InceptionV3 backbone. A linear projection layer subsequently transforms the extracted visual features into a latent space compatible with the language generation module. The choice of InceptionV3 is motivated by a few considerations. First, InceptionV3

provides a favorable trade-off between representational capacity and computational efficiency. It produces compact 2048-dimensional visual embeddings while requires substantially fewer computational resources than recent vision transformers or large vision-language models. Also, our objective was to evaluate whether a lightweight CNN backbone combined with domain-specific language models could achieve competitive performance for GI-VQA without relying on computationally expensive multimodal foundation models. Although more recent visual encoders may capture finer pathological details, InceptionV3 serves as a strong and widely adopted baseline for efficient multimodal medical image understanding.

3.1.2. Clinical Question Encoding

To capture domain-specific medical terminology and contextual information, the textual branch utilizes BioClinicalBERT as the question encoder. Given an input question Q, the tokenizer converts the question into a sequence of tokens which are processed by the transformer encoder. The contextual representation corresponding to the special classification token is extracted from the final hidden layer: $\mathbf{q} = f_{text}(Q)$, where $\mathbf{q} \in \mathbb{R}^{768}$ represents the semantic embedding of the clinical question.

BioClinicalBERT has been pre-trained on large-scale biomedical and clinical corpora, enabling improved understanding of medical concepts and GI terminology compared to general-purpose language models.

3.1.3. Multimodal Fusion and Prefix Generation

The image feature vector extracted from InceptionV3 is first projected into the BioGPT embedding space using a linear layer. Specifically, the 2048-dimensional visual representation is mapped to the decoder hidden dimension d_{BioGPT}. Similarly, the 768-dimensional BioClinicalBERT question representation is projected to the same hidden dimension using a separate linear layer. The two projected representations are then concatenated to obtain a fused multimodal vector:

$$z = [W_v v; W_q q],$$

where $W_v \in \mathbb{R}^{d_{\text{BioGPT}} \times 2048}$ and $W_q \in \mathbb{R}^{d_{\text{BioGPT}} \times 768}$ denote the visual and textual projection layers, respectively.

The concatenated representation is passed through a fusion layer that maps the $2d_{\text{BioGPT}}$-dimensional vector to $k \times d_{\text{BioGPT}}$ dimensions, followed by a $\tanh$ activation:

$$P = \tanh(W_f z + b_f),$$

where $k = 4$ is the prefix length used in our implementation. The output is reshaped into four prefix embeddings:

$$P \in \mathbb{R}^{4 \times d_{\text{BioGPT}}}.$$

These prefix embeddings encode the joint image-question context and are prepended to the BioGPT answer-token embeddings. During training, the prefix positions are masked with label value -100, so the loss is computed only over answer tokens. This allows BioGPT to generate answers conditioned on multimodal information while preserving the internal transformer architecture of the decoder.

3.1.4. Answer Generation using BioGPT

The answer generation module is based on BioGPT, a biomedical generative language model specifically designed for medical text generation tasks. During training, the generated prefix embeddings are concatenated with the answer token embeddings and provided as input to BioGPT. The model is optimized using autoregressive language modeling, where each token is predicted conditioned on the preceding tokens and the multimodal prefix representation.

The training objective is defined as:

$$\mathcal{L}_{LM} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \mathbf{P}),$$

where y_t denotes the target answer token at time step t, and $\mathbf{P}$ represents the multimodal prefix embeddings.

During inference, beam search decoding with a beam width of four is employed to improve generation quality and reduce output variability.

3.2. Implementation Details

Images were resized to 299×299 pixels, converted to RGB, and normalized using ImageNet mean and standard deviation. Questions were tokenized using the BioClinicalBERT tokenizer with a maximum length of 64 tokens. Answers were tokenized using the BioGPT tokenizer with a maximum length of 64 tokens. Padding tokens in the answer sequence were masked using a label value of -100 during training.

The model used InceptionV3 pretrained on ImageNet, BioClinicalBERT as the question encoder, and BioGPT as the answer decoder. The prefix length was set to 4. Training used a batch size of 32. During inference, answers were generated using beam search with beam width 4, a maximum generation length of 40 tokens, early stopping, and no-repeat trigram blocking with `no_repeat_ngram_size=2`.

3.3. Subtask 2

While Subtask 1 focuses solely on answer prediction, Subtask 2 requires the generation of interpretable explanations and confidence estimates alongside the predicted answer. To address this challenge, we extend our multimodal VQA framework with explainability, uncertainty quantification, and reasoning modules. The resulting system produces four outputs for every image-question pair: (i) the predicted answer, (ii) a textual explanation, (iii) a visual explanation highlighting relevant image regions, and (iv) a confidence score indicating prediction reliability.

3.3.1. Token-Level Confidence Estimation

To estimate the reliability of generated answers, we first compute token-level confidence scores directly from BioGPT's output probabilities. At each decoding step i, the model produces a probability distribution over the vocabulary:

$$P(a_i | a_{<i}, \mathbf{v}_{mm})$$

The confidence of token a_i is defined as the maximum softmax probability: $c_i = \max(P(a_i)$. The overall token confidence is computed as the average confidence across all generated tokens:

$$C_{token} = \frac{1}{N} \sum_{i=1}^{N} c_i$$

where N denotes the number of generated tokens. This measure reflects how certain the language model is during answer generation.

3.3.2. Monte Carlo Dropout Uncertainty Estimation

While token probabilities capture decoder confidence, they do not account for model uncertainty. Therefore, we employ Monte Carlo (MC) Dropout during inference. Dropout layers remain active while performing multiple stochastic forward passes through the network. Let $\{y_1, y_2, \ldots, y_M\}$ represent predictions generated from M stochastic runs. The MC confidence score is computed as:

$$C_{MC} = \frac{\max(\text{count}(y))}{M}$$

where count(y) represents the frequency of the most commonly generated prediction. Predictions exhibiting high agreement across stochastic runs are assigned higher confidence scores, whereas unstable predictions receive lower confidence values.

3.3.3. Hybrid Confidence Calibration

To obtain a more robust confidence estimate, token-level confidence and MC-Dropout confidence are combined using a weighted average: $C = 0.6C_{token} + 0.4C_{MC}$.

The weighting coefficients were selected empirically based on preliminary experiments conducted on the validation split. Several combinations were evaluated, including equal weighting (0.5/0.5) and token-dominated or uncertainty-dominated settings. We observed that assigning a slightly larger weight to the token-level confidence produced more stable confidence estimates while still allowing Monte Carlo Dropout to capture predictive uncertainty.

The resulting confidence score is normalized to the range $[0, 1]$ and included in the final submission. This strategy allows the confidence estimate to reflect both decoder certainty and model uncertainty simultaneously.

3.3.4. Visual Explanation Generation using Grad-CAM

To provide visual evidence supporting the generated answer, we employ Grad-CAM. It is applied to the final convolutional block (`Mixed_7c`) of the InceptionV3 encoder. Let A^k denote the k^{th} feature map produced by the selected layer. For a target prediction score y, channel importance weights are computed as:

$$\alpha_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y}{\partial A^k_{ij}}$$

where Z denotes the spatial dimensions of the feature map. The Grad-CAM localization map is subsequently obtained as

$$L_{GradCAM} = ReLU\left(\sum_k \alpha_k A^k\right)$$

The resulting heatmap is normalized and superimposed on the original endoscopic image to highlight diagnostically relevant regions contributing to the generated answer. For every image-question pair, the generated heatmap is stored as a visual explanation and included in the challenge submission.

3.3.5. Image-Question Semantic Alignment

To quantify the degree of correspondence between the visual content and the clinical question, cosine similarity is computed between projected image and question embeddings. The alignment score is defined as:

$$S_{align} = \frac{\mathbf{v}_{img} \cdot \mathbf{v}_{txt}}{\|\mathbf{v}_{img}\| \|\mathbf{v}_{txt}\|}$$

Higher values indicate stronger semantic consistency between the image and the associated question. This alignment score is subsequently incorporated into explanation generation.

3.3.6. Attention Localization Analysis

While Grad-CAM highlights important image regions, we additionally quantify the concentration of attention using the Gini coefficient. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ denote normalized Grad-CAM activation values sorted in ascending order. The Gini coefficient is computed as:

$$G = \frac{2\sum_{i=1}^{n} i x_i}{n \sum_{i=1}^{n} x_i} - \frac{n+1}{n}$$

where n represents the number of activation values. A high Gini coefficient indicates that the model concentrates its attention on a small set of localized regions, whereas lower values correspond to diffuse attention patterns. This metric serves as an additional measure of explanation quality and localization precision.

3.3.7. Textual Explanation Generation

To provide human-interpretable reasoning, a dedicated explanation generation module is constructed using BioGPT. For each prediction, a structured prompt is created containing:

- Clinical question
- Generated answer
- Mean token confidence
- Highest-confidence generated tokens
- Lowest-confidence generated tokens
- Image-question alignment score
- Grad-CAM localization statistics
- Attention concentration (Gini coefficient)
- Confidence calibration score

These reasoning signals are combined and provided as contextual input to BioGPT, which generates a free-text explanation describing the evidence supporting the prediction. The explanation generation process encourages the language model to justify predictions using both visual and textual evidence extracted from the multimodal framework.

4. Results and Discussion

This section presents the results obtained for Subtask 1 across different evaluation metrics mentioned in the overview paper of the task [18]. For Subtask 2 the results will be presented at the conference and will be published on Github[1] after the conference together with the leaderboard.

Our proposed VQA framework on both the public *Kvasir-VQA-test* dataset and the hidden private test set. Table 2 presents the overall results on both evaluation splits. While lexical metrics provide a measure of surface-level similarity between generated and reference answers, the LLM-based evaluation offers a more clinically meaningful assessment by measuring correctness, faithfulness, clinical relevance, clarity, and completeness.

Table 2
Overall performance on the public and private evaluation datasets.

Metric	Public	Private
BLEU-1	0.1748	0.2500
BLEU-2	0.0996	0.2000
ROUGE-1	0.6011	0.3900
ROUGE-2	0.0581	0.2500
ROUGE-L	0.5987	0.3500
METEOR	0.3744	0.4800
Correctness	7.49	6.49
Faithfulness	8.25	4.95
Clinical Relevance	7.81	6.12
Clarity	7.48	6.08
Completeness	8.40	8.22

The proposed architecture demonstrates strong generalization capabilities despite its lightweight design. Notably, the METEOR score increases from 0.3744 on the public dataset to 0.4800 on the

[1] https://github.com/simula/ImageCLEFmed-MEDVQA-GI-2026

private test set, suggesting that the generated answers maintain strong semantic alignment with the ground-truth responses even when exact wording differs.

Among the LLM-based metrics, Completeness achieves the highest score on both datasets, reaching 8.40 on the public split and 8.22 on the private split. This indicates that the model consistently addresses most components of the clinical questions. Correctness remains relatively high on the private test set (6.49), demonstrating reasonable robustness to unseen examples. The largest performance degradation is observed in Faithfulness, which decreases from 8.25 to 4.95. This suggests that although the generated answers often remain clinically relevant, the model occasionally introduces unsupported details when responding to more challenging unseen examples.

4.1. Performance Across Question Complexity

The private test set contains questions of varying complexity levels. Complexity level 1 consists of single-aspect questions, whereas levels 2 and 3 require reasoning over multiple clinical attributes simultaneously. Table 3 reports the LLM evaluation scores for each complexity level.

Table 3
LLM evaluation scores across question complexity levels on the private test set.

Complexity	Correctness	Faithfulness	Clinical	Clarity	Completeness
Level 1	7.15	3.61	5.37	5.27	8.25
Level 2	6.11	5.06	6.18	6.16	8.10
Level 3	6.14	6.34	6.90	6.89	8.32

Interestingly, the model achieves higher Faithfulness, Clinical Relevance, and Clarity scores for complexity-3 questions than for complexity-1 questions. Specifically, Faithfulness increases from 3.61 for single-aspect questions to 6.34 for triple-aspect questions, while Clinical Relevance improves from 5.37 to 6.90. This observation is somewhat counterintuitive, as more complex questions are generally expected to be more difficult. A possible explanation is that complexity-1 questions frequently require highly precise categorical answers, whereas complexity-3 questions permit more descriptive responses. Consequently, the BioGPT decoder may be better suited to generating richer explanatory outputs than concise label-like answers.

Across all complexity levels, Completeness remains consistently high, ranging from 8.10 to 8.32. This indicates that the proposed fusion mechanism effectively captures multiple semantic components within a question and enables the decoder to address them in the generated response.

4.2. Performance Across Question Categories

To better understand the strengths and weaknesses of the proposed framework, we analyze the LLM evaluation scores for individual question categories. Figure 3 shows the radar plot for each question type across each metric dimension.

The strongest performance is observed for categories related to object detection, counting, and binary recognition tasks. Instrument presence, instrument counting, text detection, and polyp removal status all achieve correctness scores above 8.5 as shown in Table 4. These tasks primarily depend on coarse visual cues that can be effectively captured by ImageNet-pretrained visual features.

In contrast, substantially lower scores are observed for landmark recognition, abnormality localization, and polyp type classification as shown in Table 5. These tasks require detailed understanding of mucosal texture, anatomical structures, and subtle pathological characteristics. The limited performance suggests that a frozen ImageNet-pretrained InceptionV3 encoder lacks the domain-specific visual representations necessary for fine-grained GI analysis.

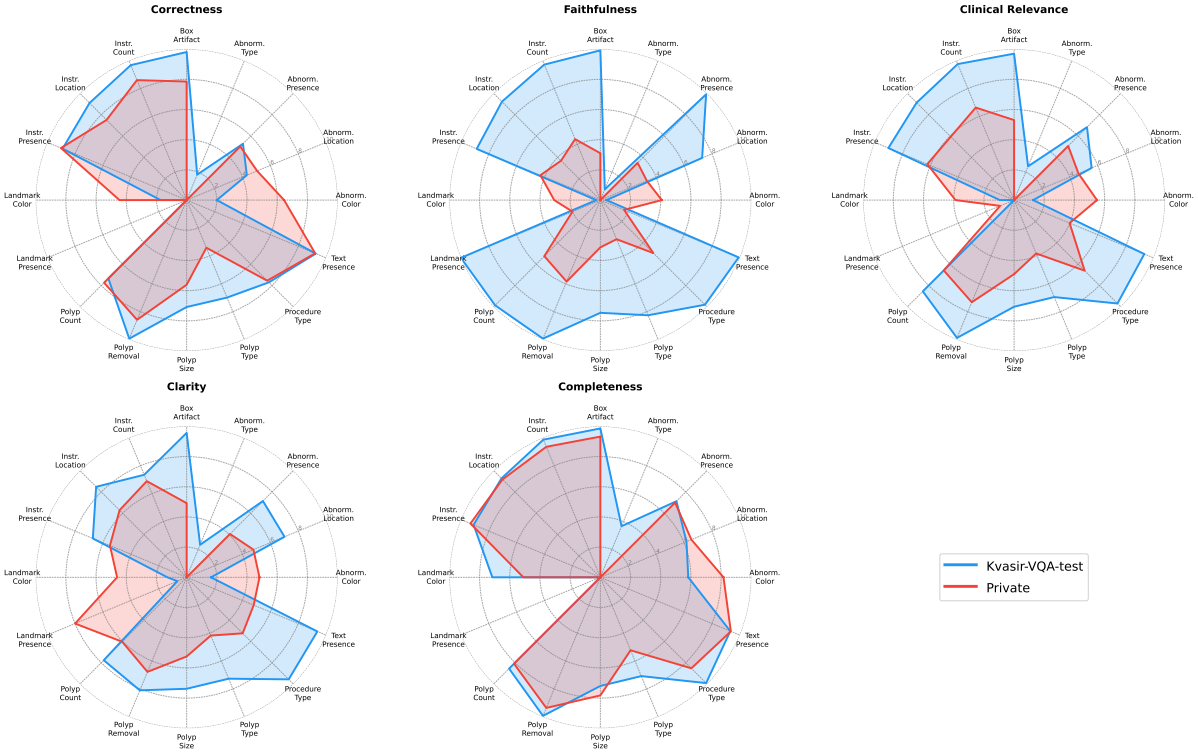

Figure 3: Radar plots for each question types across all metric dimensions

Table 4
Best-performing question categories according to Correctness on the private test set.

Question Class	Correctness
text_presence	9.21
instrument_presence	9.01
instrument_count	8.60
polyp_removal_status	8.58
box_artifact_presence	7.85

Table 5
Lowest-performing question categories according to Correctness on the private test set.

Question Class	Correctness
landmark_presence	0.00
polyp_type	3.42
landmark_color	4.45
abnormality_presence	5.03
abnormality_location	5.05

5. Conclusion and Future Work

This paper presented our participation in the ImageCLEFmed MEDVQA-GI 2026 challenge, addressing both VQA and explainable answer generation for GI endoscopy images. For Subtask 1, we developed a lightweight multimodal framework that combines InceptionV3 for visual feature extraction, Bio_ClinicalBERT for question understanding, and BioGPT for answer generation through a learned fusion mechanism. Despite its computational efficiency, the proposed system demonstrated strong performance in terms of answer completeness and clinical relevance across both public and private evaluation datasets.

For Subtask 2, we extended the framework with confidence estimation, visual grounding, semantic alignment analysis, and textual explanation generation. The integration of token-level confidence,

Monte Carlo Dropout uncertainty estimation, Grad-CAM visual explanations, and attention localization analysis enabled the model to provide interpretable predictions accompanied by reliability estimates and supporting evidence. These components contribute toward improving transparency and trustworthiness in medical AI systems.

Future research may explore endoscopy-specific visual encoders and multimodal foundation models to improve fine-grained pathological understanding. Additional directions include retrieval-augmented reasoning, counterfactual explanation generation, and uncertainty-aware decoding strategies to enhance faithfulness and reduce hallucinations. Further advances in visual grounding and explanation verification could improve the reliability and transparency of medical VQA systems. Overall, the results demonstrate the potential of lightweight, explainable multimodal architectures for clinically interpretable GI VQA.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, A. Jemal, Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer Journal for Clinicians 74 (2024) 229–263. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834. doi:https://doi.org/10.3322/caac.21834.

[2] 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, D. Johansen, H. K. Johansen, M. A. Riegler, P. Halvorsen, HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, Scientific Data 7 (2020) 1–14. doi:10.1038/s41597-020-00622-y.

[3] J. Lau, E. Lehman, P. Sharma, A dataset and exploration of models for understanding radiology reports, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2018, pp. 1–10. URL: https://aclanthology.org/D18-1001/.

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

[5] 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, in: International Conference on Learning Representations (ICLR), 2021.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763.

[7] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: a visual language model for few-shot learning, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 23716–23736.

[8] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt,

S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 19730–19742. URL: https://proceedings.mlr.press/v202/li23q.html.

[9] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, P. Rajpurkar, Med-flamingo: A multimodal medical few-shot learner, in: Proceedings of the 3rd Machine Learning for Health Symposium, PMLR, 2023, pp. 353–367. URL: https://proceedings.mlr.press/v225/moor23a.html.

[10] 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.

[11] S. Gautam, M. Riegler, P. Halvorsen, Kvasir-VQA-x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy, in: Data Engineering in Medical Imaging, Springer, 2025, pp. 53–63. doi:10.1007/978-3-032-08009-7_6.

[12] S. A. Hicks, A. Storås, P. Halvorsen, T. de Lange, M. A. Riegler, V. Thambawita, Overview of imageclefmedical 2023 – medical visual question answering for gastrointestinal tract, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[13] 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, 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.

[14] S. A. Hicks, A. Storås, P. Halvorsen, M. A. Riegler, V. Thambawita, Overview of imageclefmedical 2024 – medical visual question answering for gastrointestinal tract, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[15] 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.

[16] S. Gautam, P. Halvorsen, M. A. Riegler, V. Thambawita, S. A. Hicks, Overview of imageclefmedical 2025 – medical visual question answering for gastrointestinal tract, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[17] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, 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. Eryılmaz, 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.

[18] 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.

[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical

language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240. URL: https://doi.org/10.1093/bioinformatics/btz682. doi:10.1093/bioinformatics/btz682.
[20] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, arXiv preprint arXiv:1904.05342 (2019).
[21] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, Briefings in Bioinformatics 23 (2022) bbac409. URL: https://doi.org/10.1093/bib/bbac409. doi:10.1093/bib/bbac409.
[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Chen, W. Chen, Lora: Low-rank adaptation of large language models, in: Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR, 2021, pp. 11113–11125.
[23] Y. Brima, M. Atemkeng, Saliency-driven explainable deep learning in medical imaging: bridging visual explainability and statistical quantitative analysis, BioData Mining 17 (2024). doi:10.1186/s13040-024-00370-4.
[24] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, F. Nensa, Explainable ai in medical imaging: An overview for clinical practitioners – saliency-based xai approaches, European Journal of Radiology 162 (2023) 110787. URL: https://www.sciencedirect.com/science/article/pii/S0720048X23001018. doi:https://doi.org/10.1016/j.ejrad.2023.110787.
[25] E. H. Houssein, A. M. Gamal, E. M. G. Younis, E. Mohamed, Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review, Cluster Computing 28 (2025). URL: https://doi.org/10.1007/s10586-025-05281-5. doi:10.1007/s10586-025-05281-5.
[26] A. M. Fayyaz, S. J. Abdulkadir, N. Talpur, S. M. Al-Selwi, S. U. Hassan, E. H. Sumiea, Grad-cam (gradient-weighted class activation mapping): A systematic literature review, Computers in Biology and Medicine 198 (2025) 111200. URL: https://www.sciencedirect.com/science/article/pii/S0010482525015537. doi:https://doi.org/10.1016/j.compbiomed.2025.111200.
[27] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: M. F. Balcan, K. Q. Weinberger (Eds.), Proceedings of The 33rd International Conference on Machine Learning, volume 48 of *Proceedings of Machine Learning Research*, PMLR, New York, New York, USA, 2016, pp. 1050–1059. URL: https://proceedings.mlr.press/v48/gal16.html.
[28] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1321–1330.
[29] K. Tewari, A. Verma, S. Pal, IReL, IIT(BHU) at MEDIQA-MAGIC 2025: Tackling Multimodal Dermatology with CLIPSeg-Based Segmentation and BERT-Swin Question Answering, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS, 2025. URL: https://ceur-ws.org/Vol-4038/paper_207.pdf.
[30] K. Tewari, S. Pal, Advancing Vision and Language in GI Diagnosis: Florence2 for Question Answering and Stable Diffusion for Image Synthesis, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS, 2025. URL: https://ceur-ws.org/Vol-4038/paper_208.pdf.
[31] K. Tewari, S. Pal, X-VQA for GI Diagnostics: Multimodal Visual Question Answering with Confidence-Aware Explanations, in: Proceedings of the MediaEval 2025 Multimedia Benchmark Workshop, 2025. URL: https://2025.multimediaeval.com/paper38.pdf.
[32] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. de Lange, P. T. Schmidt, D. Johansen, Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, in: MultiMedia Modeling, Springer, 2021, pp. 218–229. URL: https://datasets.simula.no/kvasir-instrument/. doi:10.1007/978-3-030-67835-7_18.