BLIP-2-based Visual Question Answering with Multimodal Explanations for Gastrointestinal Imaging

Lakshmi Priya S[1,*,†], Dhannya S M[2,†], Moogambigai A[3,†], Nikhil Karthik S[4,†] and Pandiarajan D[5,†]

Department of Computer Science Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

Abstract

This work presents our participation in the *Medico 2025: Visual Question Answering for Gastrointestinal Imaging* challenge, addressing Subtask 1 (VQA) and Subtask 2 (Clinician-Oriented Multimodal Explanations). We employ the BLIP-2 vision–language model [1] on the Kvasir-VQA-x1 dataset for answer generation, and extend it with an explanation pipeline combining self-probing textual rationales, attention-based visual heatmaps, and confidence estimation from token probabilities. The resulting multimodal outputs enhance interpretability by linking answers with localized and reliability-aware evidence. Experimental results demonstrate that BLIP-2 serves as a competitive baseline, though further domain-specific fine-tuning and dedicated explanation modules are needed to achieve clinically faithful and trustworthy VQA performance in gastrointestinal imaging.

1. Introduction and Related Work

Endoscopic image analysis is essential for early detection of gastrointestinal (GI) diseases, and Visual Question Answering (VQA) allows clinical systems to answer diagnostic questions directly from images. The MediaEval Medico 2025 challenge [2] defines two subtasks: Subtask 1 predicts answers to GI image questions, and Subtask 2 generates multimodal explanations (textual rationales, attention-based visual localization, and confidence scores) for interpretability. Both subtasks use the Kvasir-VQA-x1 dataset [3], containing 6,500 original and 65,000 augmented images with 159,549 question–answer pairs of varying types.

Prior MedVQA research explores multimodal architectures and domain adaptation. Transformer-based GI VQA methods such as UIT-Saviors [4] benefit from image enhancement, while multimodal pretraining approaches like MMBERT [5] and graph-based methods [6] capture semantic and spatial relationships. Instruction-tuned models like PMC-VQA [7] emphasize aligned visual-text reasoning.

Building on these insights, BLIP-2 [1] is adopted as the VQA backbone for Subtask 1, fine-tuned on Kvasir-VQA-x1 [8], with an explanation pipeline for Subtask 2 combining self-probing textual rationales, attention-based visual grounding, and confidence estimation. Contributions include: (i) a BLIP-2 baseline evaluated with BLEU, ROUGE, and METEOR, (ii) a clinician-oriented explanation pipeline, and (iii) analysis of limitations of general-purpose vision–language models in the medical domain with directions for improvement.

October 25–26, 2025, Dublin, Ireland and Online

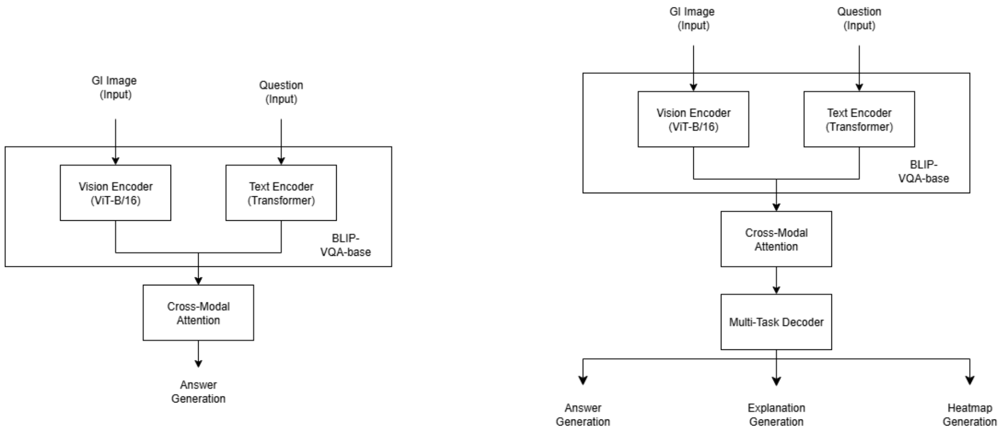
[*] Corresponding author.

[†] These authors contributed equally.

lakshmipriyas@ssn.edu.in (L. P. S); dhannyasm@ssn.edu.in (D. S. M); moogambigai2370071@ssn.edu.in (M. A); nikhilkarthik2370024@ssn.edu.in (N. K. S); pandiarajan2370062@ssn.edu.in (P. D)

2. Methodology

Our approach extends **BLIP-VQA-base** [1] into a unified multi-task framework for answer generation, textual explanation, and lesion localization (Figure 1). The architecture jointly optimizes question answering (Subtask 1) and multimodal explanation (Subtask 2) using shared encoders and task-specific heads.



(a) Subtask 1: VQA Pipeline

(b) Subtask 2: Multimodal Explanation Pipeline

Figure 1: Proposed multi-task architecture for answer generation and explanation.

2.1. VQA and Explanation Pipeline

The vision encoder (**ViT-B/16**) and transformer text encoder generate fused representations via cross-modal attention for answer decoding. Explanations follow the format *"answer: <ans> explanation: <exp>"*, with attention-rollout masks for visual grounding and token-probability averaging for confidence. The model, fine-tuned on **Kvasir-VQA-x1**, minimizes

$$\mathcal{L}_{total} = \mathcal{L}_{ans} + \lambda_{mask}\mathcal{L}_{mask}, \ \lambda_{mask} = 1.0,$$

where $\lambda_{mask} = 1.0$ balances textual and visual learning.

2.2. Implementation Details

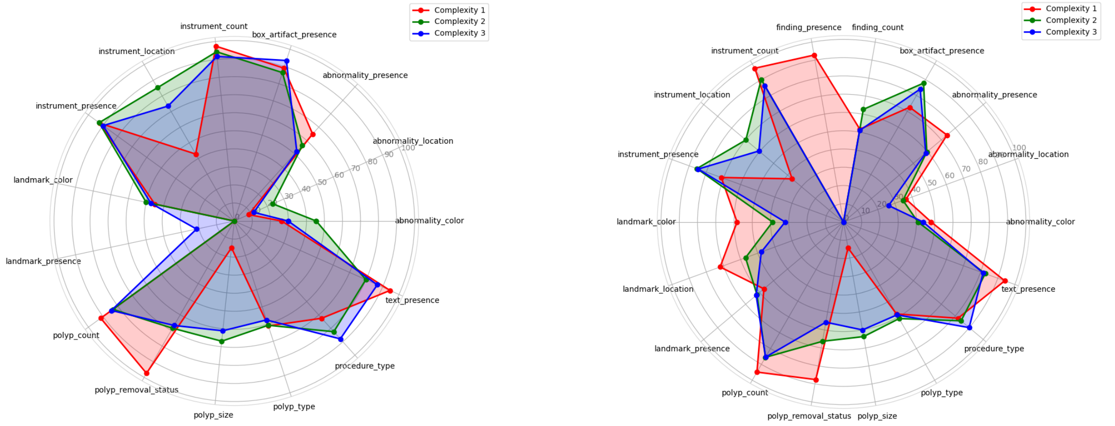
All experiments were conducted on a single GPU system running Windows 11, equipped with an Intel i7 CPU, 16 GB system RAM, and an NVIDIA RTX 4060 GPU with 16 GB VRAM. The model was implemented using PyTorch and Hugging Face Transformers, leveraging the *Trainer* API for training. **BLIP-VQA-base** was fine-tuned on the **Kvasir-VQA-x1** dataset for 5 epochs, using a batch size of 8 with gradient accumulation over 2 steps. The learning rate was set to 3×10^{-5} with the AdamW optimizer and a cosine learning rate schedule, along with a warmup ratio of 0.1. Mixed-precision training (FP16) was employed to reduce memory usage and accelerate training. Random seeds were fixed for reproducibility, and attention masks and token probabilities were logged for visual grounding evaluation. The complete implementation and code scripts are publicly available at GitHub Repository

3. Results and Analysis

This section presents the performance of the BLIP-2 baseline on **Subtask 1 (Visual Question Answering)** and **Subtask 2 (Multimodal Explanations)** from the Medico 2025 challenge.

3.1. Subtask 1: Visual Question Answering

Evaluation for Subtask 1 uses standard VQA metrics, including ROUGE [9], METEOR [10], CHRF++ [11], BLEU [12], and BERTScore [13] (Precision/Recall/F1). On the *private test set*, the BLIP-2 model achieved ROUGE-1 of 0.5431, METEOR of 0.5453, and BLEU of 0.2345. On the *public test set*, it reached ROUGE-1 of 0.5643, METEOR of 0.5712, and BLEU of 0.2556, indicating stable generalization. Radar plots in Figure 2 illustrate performance across metrics and question complexities.



(a) Private test set

(b) Public test set

Figure 2: Radar plots of evaluation metrics for Subtask 1.

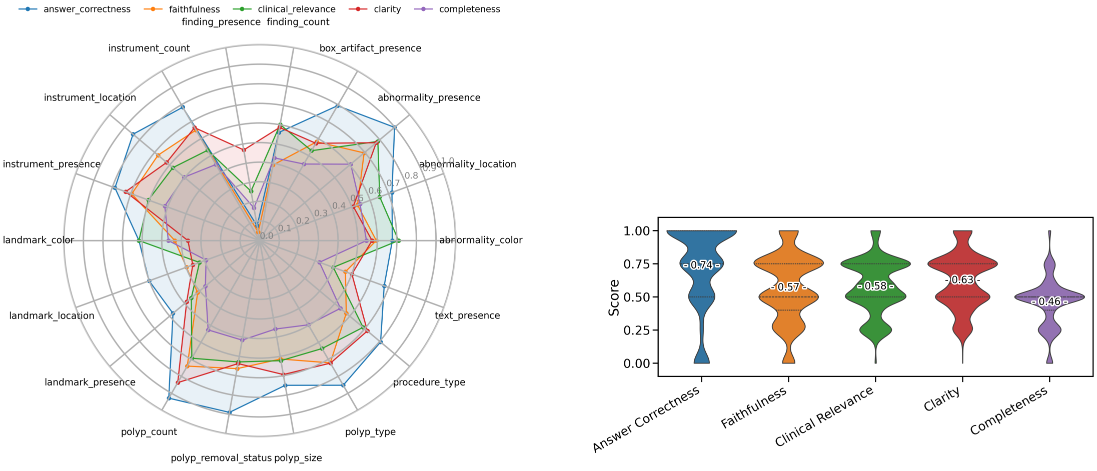
3.2. Subtask 2: Multimodal Explanations

Subtask 2 evaluates interpretability using correctness, faithfulness, clinical relevance, clarity, and completeness. The BLIP-VQA-base model achieves mean scores of 0.7367 (correctness), 0.5706 (faithfulness), and 0.4613 (completeness). Figure 3 shows performance variation by question type, with higher accuracy for polyp-related queries and lower reliability for subtle landmarks. Subtask 2 was trained exclusively on the final Subtask 1 model; no ablation experiments were conducted for explanation components.

4. Ablation Study

To investigate the effect of backbone architectures on Subtask 1 (Visual Question Answering), several model combinations were evaluated on Kvasir-VQA-x1. Table 1 presents BLEU scores for each configuration.

The results show that generic transformer backbones (BERT + ViT) achieve moderate performance, while MedicalBERT provides a slight improvement due to domain-specific pretraining.



(a) Scores by question type

(b) Overall score distribution

Figure 3: Subtask 2 performance across explanation metrics.

Table 1
Ablation results for Subtask 1: BLEU scores across different model combinations.

Model Configuration	BLEU Score
BERT + Vision Transformer	0.12
MedicalBERT + Vision Transformer	0.15
BLIP	0.05
BLIP-VQA-base (final model)	0.2345

BLIP alone underperforms, indicating the importance of joint vision-language pretraining. The BLIP-VQA-base model yields the highest BLEU score and serves as the backbone for Subtask 2.

5. Discussion and Future Work

The BLIP-2 baseline demonstrates strong semantic understanding (BLEU = 0.2528, ROUGE-1 = 0.5635, METEOR = 0.5679) but limited lexical precision. Subtask 2 results (correctness = 0.7367, faithfulness = 0.5706) suggest solid grounding for visually salient tasks (e.g., polyp counts), yet weaker reasoning for abstract or spatial queries. Occasional mismatches between answers and explanations highlight limitations in multimodal alignment. BLIP-2 shows reasonable understanding but limited precision and abstract reasoning; future work should focus on GI-specific fine-tuning, improved attention grounding, and enhanced text–vision synchronization.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] J. Li, D. Li, S. Savarese, S. C. H. Hoi, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, in: Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR, 2023. URL: https://proceedings.mlr.press/v202/li23k.html.

[2] S. Gautam, V. Thambawita, M. Riegler, P. Halvorsen, S. Hicks, Medico 2025: Visual question answering for gastrointestinal imaging, arXiv preprint arXiv:2508.10869 (2025).

[3] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-VQA-x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy, arXiv (2025). doi:10.48550/arXiv.2506.09958. arXiv:2506.09958.

[4] T. M. Thai, B. T. Nguyen, B. T. Nguyen, Q. D. Le, A. T. Nguyen, N. D. Le, Uit-saviors at medvqa-gi 2023: Improving multimodal learning with image enhancement, arXiv preprint arXiv:2307.02783 (2023).

[5] Y. Khare, S. K. Tamang, V. Gudivada, Mmbert: Multimodal bert pretraining for improved medical vqa, arXiv preprint arXiv:2104.01394 (2021).

[6] H. Hu, Y. Wang, et al., Interpretable medical image visual question answering via multi-modal relationship graph learning, Medical Image Analysis 97 (2024) 103159. doi:10.1016/j.media.2024.103159.

[7] X. Zhang, Y. Yang, et al., Pmc-vqa: Visual instruction tuning for medical visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. URL: https://xiaoman-zhang.github.io/PMC-VQA/.

[8] 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.

[9] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Association for Computational Linguistics, 2004, pp. 74–81.

[10] S. Banerjee, A. Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, 2005, pp. 65–72.

[11] M. Popović, CHRF++: Words Helping Character n-grams Improve Machine Translation Evaluation, in: Proceedings of the Second Conference on Machine Translation (WMT 2017), Association for Computational Linguistics, 2017, pp. 612–618.

[12] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2002, pp. 311–318. doi:10.3115/1073083.1073135.

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, in: Proceedings of the International Conference on Learning Representations (ICLR), 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.