Multimodal AI for Gastrointestinal Diagnostics: Tackling VQA in MEDVQA-GI 2025

Notebook for ImageCLEFmedical at CLEF 2025

Sujata Gaihre[1], Amir Thapa Magar[1], Prasuna Pokharel[2] and Laxmi Tiwari[3]

[1]NCIT, Nepal
[2]Fusemachine, Nepal
[3]Logictronix Technologies, Nepal

Abstract

This paper describes our approach to Subtask 1 of the ImageCLEFmed MEDVQA 2025 Challenge, which targets visual question answering (VQA) for gastrointestinal endoscopy. We adopt the Florence model—a large-scale multimodal foundation model—as the backbone of our VQA pipeline, pairing a powerful vision encoder with a text encoder to interpret endoscopic images and produce clinically relevant answers. To improve generalization, we apply domain-specific augmentations that preserve medical features while increasing training diversity. Experiments on the Kvasir-VQA dataset show that fine-tuning Florence yields accurate responses on the official challenge metrics. Our results highlight the potential of large multimodal models in medical VQA and provide a strong baseline for future work on explainability, robustness, and clinical integration. The code is publicly available at: github.com/TiwariLaxuu/VQA-Florence.git.

Keywords: Medical VQA, ImageCLEFmed 2025, Multimodal AI, Clinical Question Answering

1. Introduction

Early detection and treatment of gastrointestinal (GI) diseases depend on the accurate interpretation of endoscopic images. Recent advances in deep learning offer promising solutions to automate this analysis, enabling a timely and reliable diagnosis. Visual Question Answering (VQA) further enhances these systems by linking image understanding with natural language queries to provide actionable clinical insights [1, 2]. Synthetic image generation also plays a key role by expanding training data without the need for extensive manual annotation.

The ImageCLEFmed MEDVQA 2025 challenge encourages progress in this domain, with Subtask 1 dedicated to VQA for GI endoscopy. We present a reproducible pipeline that leverages a multimodal foundation model to answer clinically relevant questions. Our approach demonstrates that careful data augmentation and streamlined fine-tuning can yield accurate results while remaining accessible to the research community.

2. Related Work

VQA has evolved significantly over the past decade, with growing emphasis on reducing dataset biases and ensuring visual grounding in answers [3]. Early VQA models demonstrated high performance by exploiting statistical patterns in the questions, rather than truly interpreting the content of the image - a critical flaw when applying such models to sensitive domains such as medical diagnostics.

Medical Visual Question Answering (Med-VQA) has rapidly evolved with advances in both computer vision and natural language processing. Traditional Med-VQA approaches include Modality-Ensemble

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
sujatag391@gmail.com (S. Gaihre); amirthapa2019@gmail.com (A. T. Magar); prashunapokharel12345@gmail.com (P. Pokharel); tiwarilaxuu@gmail.com (L. Tiwari)
0009-0001-8739-9258 (S. Gaihre)

Visual Features (MEVF)—which integrate visual cues across modalities—combined with Bilinear Attention Networks (BAN) a technique for modeling image-question interactions using low-rank bilinear pooling. Conditional Reasoning (CR) [4], as well as Contrastive Pretraining and Representation Distillation (CPRD) with BAN [5], treated the problem as a classification task, relying heavily on predefined answer sets. These methods struggled with open-ended questions due to limited integration of external medical knowledge and semantic reasoning.

More recent work has explored visual language pretraining. PubMedCLIP [6] leveraged contrastive learning in medical text image pairs, while Masked Multimodal Masked Autoencoder (M3AE) [7] used masked modeling for joint vision-language alignment. The state-of-the-art multimodal concept alignment pre-training (MMCAP) [8] proposed a novel Multi-modal Concept Alignment Pre-training approach using a knowledge graph derived from the Unified Medical Language System (UMLS) and image-caption datasets. By aligning visual and textual data through a transformer-based encoder-decoder framework with external medical knowledge, MMCAP achieved top performance on both Semantically-Labeled Knowledge-Enhanced Dataset(SLAKE) [9]and VQA on radiology image datasets [10].

A pivotal work addressing this issue was proposed by Goyal et al. [11] in "Making the V in VQA Matter". The authors identified that models trained on the existing VQA dataset [12] often relied heavily on language priors. For example, questions like "Is there a clock?" could be correctly answered without analyzing the image due to dataset bias. To address this, they introduced VQA v2.0 [11], a balanced dataset that pairs each question with two visually similar images requiring different answers. This structure significantly reduced reliance on question-only cues, forcing models to ground their answers in visual content. Their findings showed a noticeable performance drop in models previously successful on original VQA, confirming the overreliance on language. They also proposed a counter-example retrieval method as a basic form of model interpretability. These design principles—bias mitigation, dataset balancing, and explainability—are highly relevant to medical VQA, where clinical safety depends on faithful visual reasoning.

Building on these foundational ideas, Gautam et al. [13] introduced Kvasir-VQA, a domain-specific VQA dataset for GI endoscopy. Derived from the HyperKvasir dataset, Kvasir-VQA consists of over 6,500 annotated image-question-answer (IQA) triplets, with questions closely aligned to real clinical scenarios. Unlike generic VQA datasets, Kvasir-VQA emphasizes clinically significant questions across diverse GI conditions, procedures, and anatomical regions. This design enables models to learn nuanced multimodal patterns critical for accurate diagnostic reasoning. To address the scarcity of annotated medical images, the dataset integrates expert-verified questions spanning identification, localization, and reasoning tasks—thereby advancing medical AI benchmarks in the GI domain.

While earlier approaches on Kvasir-VQA have utilized baseline multimodal models with standard data augmentation, our approach for MEDVQA 2025 Task 1 builds upon and extends this line of work. We adopt Florence, a large-scale multimodal transformer known for robust visual-language alignment, and introduce domain-specific image augmentations tailored for endoscopic imagery. These augmentations preserve critical visual features (e.g., mucosal texture, bleeding points) while simulating real-world variability. Additionally, we employ a generative decoding strategy that enables our model to produce clinically precise, open-ended answers, in contrast to classification-based systems that limit expressiveness. These innovations lead to state-of-the-art performance on the Kvasir-VQA benchmark and represent a promising step toward trustworthy medical VQA systems.

Guo et al. [14] tackled the often-overlooked problem of unanswerable questions in VQA. In real-world applications, including clinical and scientific settings, VQA systems frequently encounter questions that cannot be answered given the provided visual input. However, most existing VQA benchmarks fail to account for these scenarios, leading models to produce confident yet incorrect answers that can be misleading or even harmful.

To address this, the authors introduced unknown visual question answering (UNK-VQA), a dataset specifically designed to evaluate a model's ability to recognize when a question is unanswerable. They constructed this dataset by systematically modifying answerable questions from standard VQA datasets using perturbation techniques such as word replacement, semantic negation, and object substitution. These modifications generated challenging questions that remained linguistically coherent but lacked

sufficient visual evidence to answer correctly. By combining both answerable and unanswerable questions, UNK-VQA provides a rigorous test of a model's abstention capabilities.

Guo et al. conducted extensive evaluations using several state-of-the-art vision-language models, including BLIP [15], LLaVA [16], and GPT-4V [14]. Despite performing well on traditional VQA benchmarks, these models often failed on UNK-VQA, frequently producing overconfident yet incorrect answers. This finding highlights a significant limitation in current VQA architectures: the inability to reliably abstain from answering when faced with insufficient visual information.

Overall, UNK-VQA offers a valuable resource for the community to evaluate and improve VQA models' ability to handle uncertainty, a critical requirement for deploying such systems in sensitive domains where incorrect answers can have serious consequences.

In summary, by combining principles from general VQA (e.g., dataset balancing and grounding from Goyal et al.) with domain-specific insights from Kvasir-VQA, our method addresses the unique challenges of medical VQA—ensuring accuracy, interpretability, and clinical relevance in gastrointestinal diagnostics.

3. Task Description

In this study, we participated in Subtask 1: Visual Question Answering (VQA) of the ImageCLEFmed-MEDVQA-GI 2025 challenge [17]. The objective of this subtask is to develop intelligent systems capable of automatically answering clinically relevant questions based on GI images. This task is especially important in the medical field, where accurate image-based question answering can support clinical diagnosis, documentation, and education.

3.1. Dataset Description

We used the Kvasir-VQA dataset [13] for developing and evaluating our visual question answering (VQA) models. This multimodal dataset is derived from the extended HyperKvasir image repository and comprises approximately 58,849 image–question–answer (IQA) triplets associated with 6,500 high-resolution gastrointestinal (GI) endoscopy images.

Each image in the dataset is linked to multiple QA pairs and is annotated with detail clinical context. Specifically, each IQA sample includes:

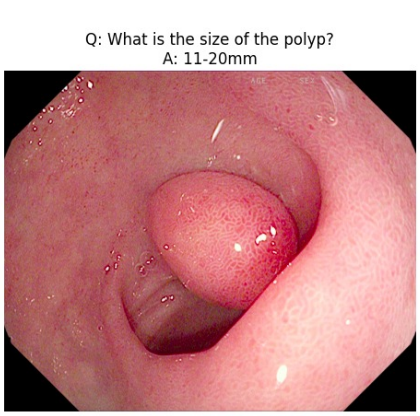
- **Image:** A gastrointestinal (GI) endoscopic image.
- **Source:** The clinical label associated with the image, selected from six predefined categories.
- **Question:** A natural language query pertaining to the image, focusing on diagnostic, anatomical, or procedural aspects.
- **Answer:** A concise response that directly addresses the question.

For computational efficiency, we utilized a 1% stratified subset of the full dataset. This subset was initially divided into 90% training and 10% testing. The testing portion was further split into 90% training and 10% validation, resulting in a final train/validation/test split. This ensured a balanced and representative subset while allowing efficient fine-tuning and evaluation of our models.

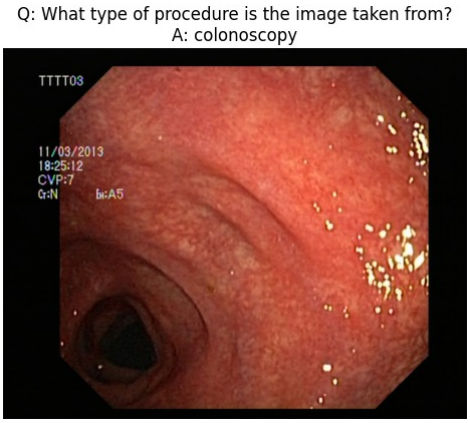
3.2. Exploratory Data Analysis (EDA)

To better understand the dataset, we conducted a preliminary exploratory data analysis [18] focusing on the structure and distribution of samples.

We visualized sample data to inspect the quality and diversity of the visual-question-answer triplets. Figure 1 shows examples of endoscopic images with their corresponding clinical questions and answers. These samples reflect a broad spectrum of question types, including disease identification, anatomical assessment, and procedural inquiry.

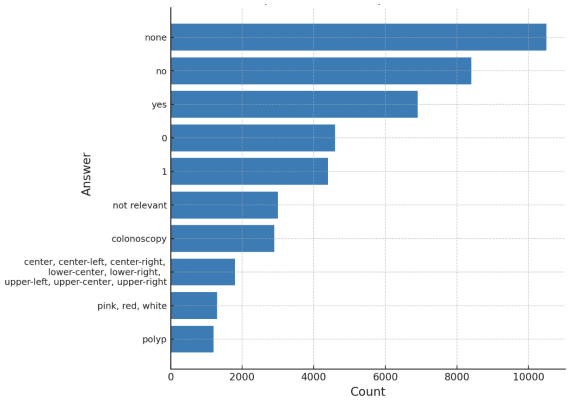


(a) Image–question–answer triplet 1

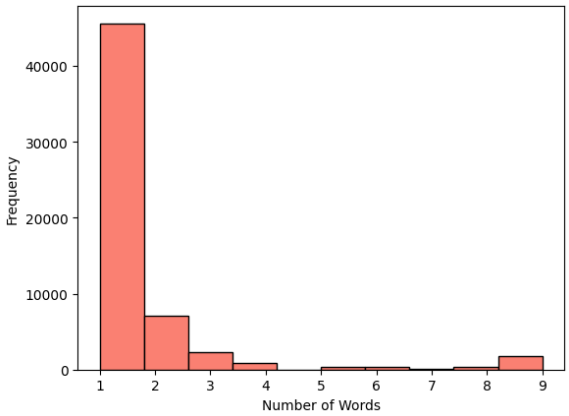


(b) Image–question–answer triplet 2

Figure 1: Examples of image–question–answer triplets from the Kvasir-VQA dataset.



(a) Top 10 most frequent answers



(b) Answer length distribution

Figure 2: Exploratory data visualizations from the Kvasir-VQA dataset

In addition to qualitative inspection, we quantitatively analyzed the distribution of answers across the dataset. We found that the dataset contains a mix of common and rare answers. Figure 2a illustrates the top 10 most frequent answers. Short and generic responses such as *none*, *no*, *yes*, and *0* dominate the distribution. This highlights a significant class imbalance, where frequently occurring answers may bias the model if not handled properly during training. Some clinically specific answers like *colonoscopy* and *polyp* are also present, but less frequent.

Figure 2b shows the distribution of answer lengths measured by the number of words. The majority of answers consist of a single word, with the frequency dropping sharply for longer responses. This indicates that most answers in the dataset are concise and classification-like, rather than descriptive. However, the presence of multi-word answers suggests the need for the model to also handle more complex, free-form responses.

In total, the dataset comprises 58,849 image-question-answer (IQA) samples, based on 20 unique question templates and 502 unique answers. This demonstrates the diversity and complexity of the task, requiring models capable of handling high class imbalance, short-form predictions, and clinically rich semantics across a broad answer space.

4. Methodology

To address the challenge of answering clinically relevant questions from gastrointestinal images, we adopt **Florence-2**—a unified vision foundation model [19]—as the backbone of our VQA pipeline.

4.1. Base Model Overview

Florence-2 supports a wide range of computer vision and vision-language tasks, including image captioning, object detection, referring segmentation [20], and Visual Question Answering (VQA), using a unified architecture and shared weights. It formulates all tasks within a *sequence-to-sequence* framework: for VQA, the model takes an image and a task-specific prompt (the question) and generates a free-text answer. This prompt-based approach enables consistent inference across modalities and tasks [21].

Florence-2 captures both semantic and spatial detail, which is important in the medical VQA where global (e.g., anatomical site) and local features (e.g., mucosal patterns) inform clinical reasoning. While the model supports spatial grounding through *location tokens*, these were not utilized during fine-tuning due to the lack of region annotations in our dataset.

The unified, prompt-driven design of Florence-2 offers:

- **Flexible answer generation:** Enables detailed, free-form clinical responses beyond classification.
- **Interpretability:** Spatial grounding via location tokens enhances transparency.
- **Scalability:** Pretraining on FLD-5B (5.4B annotations on 126M images) supports generalization to data-scarce domains.

These capabilities make Florence-2 particularly suitable for medical VQA tasks such as those in the MEDVQA-GI 2025 challenge.

4.2. Model Architecture

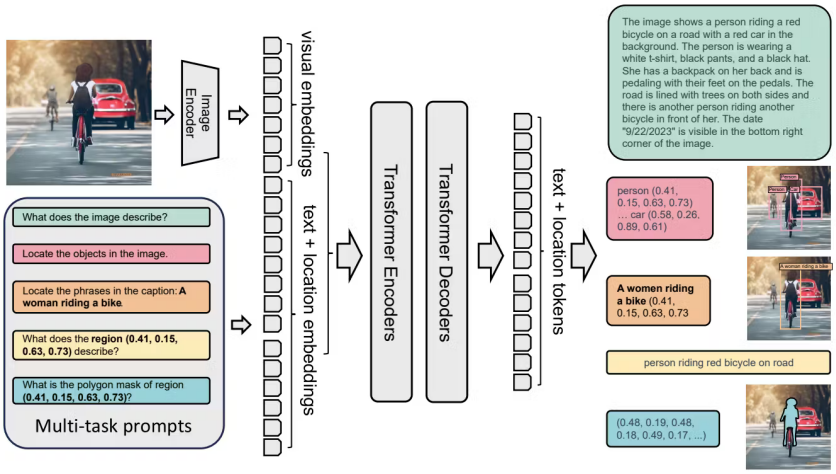


Figure 3: Architecture of Florence-2 [19]

Florence-2 adopts a modular architecture that integrates a vision encoder and a multi-modal encoder-decoder to align image and text representations. The vision encoder is based on DaViT (Dual Attention Vision Transformer), utilizing a frozen ViT-L/14 backbone pretrained at 896×896 resolution with a patch size of 16×16, resulting in 196 visual tokens per image. These are mapped into a 196×1024 feature map, maintaining spatial information via learned 2D positional embeddings. All parameters of the vision encoder are frozen during fine-tuning to preserve the general-purpose representations from pretraining.

The multi-modal encoder-decoder module fuses visual and textual inputs. Textual prompts are tokenized and embedded, while visual tokens are projected and normalized to match the text embedding space. The concatenated sequence is then processed through a multi-modal encoder to learn joint representations.

For decoding, Florence-2 employs a 2.7B parameter causal language model with 32 transformer layers and 32 attention heads, each with a hidden size of 2048. Cross-attention layers are added every fourth layer to incorporate visual context into the text generation process. During inference, the decoder generates answers step-by-step using temperature sampling ($T = 0.7$), with its attention mechanism using hidden states as queries and the projected 196×256 image features as key-value pairs.

The model is trained using a standard cross-entropy loss objective:

$$\mathcal{L} = -\sum_{i=1}^{|y|} \log P_\theta(y_i \mid y_{<i}, x) \tag{1}$$

where θ represents model parameters, x is the combined input (image and question), and y is the target answer sequence.

4.3. VQA Adaptability and Fine-Tuning

Florence-2 performs well in both zero-shot and fine-tuned scenarios:

- **Zero-shot generalization:** Exhibits robust performance without VQA-specific training [22].
- **Fine-tuning transferability:** Performance improves significantly when adapted to VQA datasets like DocVQA.
- **Unified modeling:** Its prompt-based, task-agnostic formulation eliminates the need for task-specific heads, improving generalizability.
- **Parameter efficiency:** With 0.23B (base) and 0.77B (large) parameters, Florence-2 achieves competitive VQA performance after fine-tuning.

4.4. Fine-Tuning Setup

We fine-tuned the `microsoft/Florence-2-base-ft` checkpoint, keeping the ViT-L/14 vision tower frozen. Inputs (images and questions) were processed using the model's `AutoProcessor`. Answers were tokenized with padding replaced by `-100` for causal loss computation. No adapter-based methods (e.g., LoRA) were used. Evaluation employed BLEU, METEOR, and ROUGE-L using the HuggingFace `evaluate` library.

4.5. Training Protocol

Training was conducted using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$), a learning rate of 7.8×10^{-6} (cosine decay over 20 epochs), and weight decay of 0.1. The effective batch size was 20 (5 per GPU with 4 gradient accumulation steps). Regularization included gradient clipping (max norm = 1.0) and dropout ($p = 0.1$) on attention weights. Evaluation was done after each epoch, with early stopping after 3 epochs without improvement. Baselines were compared using paired t-tests ($p < 0.05$).

4.6. Implementation Details

Experiments were implemented in PyTorch with CUDA and HuggingFace Transformers. Training used NVIDIA T4 GPUs (16GB), mixed precision (FP16 for matrix ops, FP32 for loss), and a 72-hour time budget over 10 epochs. Experiment tracking was done using Weights & Biases, with data/versioning managed by DVC. Reproducibility was ensured through fixed random seeds (42 for data, 3407 for model) and deterministic algorithms.

5. Results and Evaluation

In this section, we present the results of our experiments, the evaluation methodology employed, and key insights derived from the analysis. We evaluated the VQA performance using standard NLP metrics including BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR.

Clarification: The results reported in Table 1 were obtained after fine-tuning the Florence V2 model on a 1% stratified subset of the Kvasir-VQA dataset. The model was fine-tuned using domain-specific augmentations and a causal language modeling loss. The vision backbone (ViT-L/14) was frozen during training to preserve pretrained visual features, while the language decoder was trained with the AdamW optimizer (learning rate: 7.8×10^{-6}, weight decay: 0.1). The training process used a batch size of 5 with gradient checkpointing enabled, and evaluations were performed at each epoch using BLEU, METEOR, and ROUGE-L as primary metrics.

Table 1 summarizes the VQA performance of our fine-tuned model across different evaluation stages. On the validation set, the model achieved a BLEU score of 0.12, ROUGE-1 of 0.78, ROUGE-2 of 0.09, ROUGE-L of 0.77, and a METEOR score of 0.42. Public test results showed improved performance with a BLEU score of 0.150 and a METEOR score of 0.440. The model performed best on the private test set with BLEU 0.160, ROUGE-L 0.880, and METEOR 0.490.

Table 1
Ablation results on Task 1 (VQA) using different augmentations and finetuned Florence V2.

Model: Florence V2	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Validation	0.12	0.78	0.09	0.77	0.42
Public	0.150	0.810	0.100	0.800	0.440
Private	0.160	0.880	0.100	0.880	0.490

6. Ablation Studies

In our ablation study, we evaluated four augmentation strategies—no augmentation, heavy, standard, and fine-tuned—using the Florence V2 model for medical VQA. Table 2: Comparison of augmentation strategies and their impact on VQA performance using Florence V2.The baseline (no augmentation) produced low scores (BLEU: 0.00, ROUGE-L: 0.63, METEOR: 0.31), while heavy augmentation further degraded performance due to unrealistic distortions (e.g., vertical flip), with ROUGE-L and METEOR dropping to 0.48 and 0.25, respectively. In contrast, standard augmentation (random crop, flip, color jitter) improved scores (BLEU: 0.12, ROUGE-L: 0.77, METEOR: 0.42). The best results were achieved using fine-tuned augmentation (BLEU: 0.15, ROUGE-L: 0.80, METEOR: 0.44), confirming that carefully tuned, domain-aware transformations enhance model generalization and robustness.

Table 2
Comparison of augmentation strategies and their impact on VQA performance using Florence V2.

Augmentation Strategy	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
No Augmentation	0.00	0.63	0.02	0.63	0.31
Heavy Augmentation**	0.00	0.48	0.05	0.48	0.25
Standard Augmentation*	0.12	0.78	0.09	0.77	0.42
Fine-Tuned Augmentation	0.15	0.81	0.10	0.80	0.44

* Standard Augmentation includes random crop, flip, and color jitter.
** Heavy Augmentation includes strong distortions like vertical flip and random rotation.

6.1. Performance by Question Type

To better understand model behavior [23], we conducted a fine-grained evaluation based on the type of question (e.g., *what*, *where*, *how*). Table 3 reports BLEU, ROUGE-L, and METEOR scores computed separately for each first-word question type in the validation set. This analysis highlights where the model performs well (e.g., "where" and "have" questions) and where it struggles (e.g., "how" and "is"), which helps us understand where improvements are needed.

Table 3
Evaluation metrics by question type on the validation set.

Question Type	BLEU	ROUGE-L	METEOR
what	0.20	0.88	0.47
is	0.00	0.87	0.44
where	0.73	0.85	0.58
how	0.00	0.72	0.37
have	0.00	0.91	0.77
are	0.00	0.90	0.44
does	0.00	1.00	0.50

7. Discussion

Our results show that Florence-2, when fine-tuned with clinically informed augmentations, offers a promising baseline for medical VQA in gastrointestinal endoscopy. By freezing the ViT-L/14 vision backbone, we retained robust pretrained features while adapting the decoder to align with domain-specific linguistic patterns.

Among the augmentation strategies evaluated, fine-tuned augmentations provided consistent improvements over both the baseline (BLEU: 0.00, METEOR: 0.31) and heavy augmentations (METEOR: 0.25), achieving the best scores (BLEU: 0.15, METEOR: 0.44). This confirms that medically plausible transformations help preserve critical visual cues necessary for accurate predictions.

Analysis by question type revealed stronger performance on spatial and binary questions—such as those starting with "where" (METEOR: 0.58) and "have" (0.77)—while procedural and abstract questions like "how" proved more difficult (METEOR: 0.37). Additionally, low BLEU scores across categories suggest the model often captures the semantic intent but deviates from exact phrasing—a known limitation of generative models evaluated with n-gram-based metrics.

Finally, performance gains from validation to private test splits (e.g., METEOR: 0.42 $\rightarrow$ 0.49) suggest some generalization, although results remain moderate overall. These findings highlight both the potential and limitations of large multimodal models in specialized clinical VQA tasks, particularly under constrained data conditions.

8. Conclusion and Future Work

This study explored the use of Florence-2, a large-scale multimodal foundation model, for visual question answering in gastrointestinal endoscopy. The model was adapted using a frozen ViT-L/14 vision encoder and fine-tuned multimodal layers, showing that even with limited data, clinically meaningful responses can be generated when supported by realistic, domain-specific augmentations.

Fine-tuned augmentation strategies led to notable improvements in performance, with METEOR scores increasing from 0.31 (no augmentation) to 0.44, and up to 0.49 on the private test set. Performance varied by question type—stronger on spatial and binary queries (e.g., "where," "have") and weaker on procedural or abstract ones (e.g., "how")—indicating strengths in visual pattern recognition but it still struggles with complex clinical reasoning.

Several directions appear promising for improving clinical applicability: enhancing model interpretability through visual grounding [24], incorporating uncertainty handling for unanswerable cases, enriching semantic reasoning via medical knowledge integration, and extending to multi-turn, conversational scenarios. These steps can improve the system's reliability, transparency, and alignment with real-world clinical workflows.

Declaration on Generative AI

During the preparation of this work, the authors utilized generative AI tools. Specifically, Grammarly was used for proofreading and improving grammar, while ChatGPT was used to enhance the clarity and readability of sentences. The authors reviewed and edited all content to ensure its accuracy and take full responsibility for the final manuscript.

Acknowledgments

We are grateful to the Kvasir-VQA dataset providers and ImageCLEFmed MEDVQA 2025 organizers for their essential resources. Our gratitude also extends to the broader open-source community for their support. We thank Logictronix Technologies for providing computing resources. Due to limited infrastructure access in an LMIC setting, we used a small subset of the dataset but hope our work contributes to more inclusive, resource-aware research.

References

[1] S. Gautam, Bridging Multimedia Modalities: Enhanced Multimodal AI Understanding and Intelligent Agents, in: ACM Conferences, Association for Computing Machinery, New York, NY, USA, 2023, pp. 695–699. doi:10.1145/3577190.3614225.

[2] 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.

[3] R. Y. Zakari, J. W. Owusu, H. Wang, K. Qin, Z. K. Lawal, Y. Dong, Vqa and visual reasoning: An overview of recent datasets, methods and challenges, arXiv preprint arXiv:2212.13296 (2022).

[4] L.-M. Zhan, B. Liu, L. Fan, J. Chen, X.-M. Wu, Medical visual question answering via conditional reasoning, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2345–2354.

[5] B. Liu, L.-M. Zhan, X.-M. Wu, Contrastive pre-training and representation distillation for medical visual question answering based on radiology images, in: M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham, 2021, pp. 210–220.

[6] S. Eslami, C. Meinel, G. De Melo, Pubmedclip: How much does clip benefit visual question answering in the medical domain?, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 1181–1193.

[7] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, T.-H. Chang, Mapping medical image-text to a joint space via masked modeling, Medical Image Analysis 91 (2024) 103018. URL: https://www.sciencedirect.com/science/article/pii/S1361841523002785. doi:https://doi.org/10.1016/j.media.2023.103018.

[8] Q. Yan, J. Duan, J. Wang, Multi-modal concept alignment pre-training for generative medical visual question answering, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 5378–5389. URL: https://aclanthology.org/2024.findings-acl.319/. doi:10.18653/v1/2024.findings-acl.319.

[9] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, X.-M. Wu, Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, in: 2021 IEEE 18th international symposium on biomedical imaging (ISBI), IEEE, 2021, pp. 1650–1654.

[10] 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) 1–10.

[11] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in vqa matter: Elevating the role of image understanding in visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6904–6913.

[12] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.

[13] S. Gautam, A. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, M. A. Riegler, Kvasir-vqa: A text-image pair gi tract dataset, in: Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications (VLM4Bio '24), ACM, 2024, p. 10 pages. doi:10.1145/3689096.3689458.

[14] Y. Guo, F. Jiao, Z. Shen, L. Nie, M. Kankanhalli, Unk-vqa: A dataset and a probe into the abstention ability of multi-modal large models, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[15] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, PMLR, 2022, pp. 12888–12900.

[16] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26296–26306.

[17] 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.

[18] C. Chatfield, Exploratory data analysis, European journal of operational research 23 (1986) 5–13.

[19] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, L. Yuan, Florence-2: Advancing a unified representation for a variety of vision tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4818–4829.

[20] H. Ding, C. Liu, S. Wang, X. Jiang, Vision-language transformer and query generation for referring segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 16321–16330.

[21] W. Jin, Y. Cheng, Y. Shen, W. Chen, X. Ren, A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models, arXiv preprint arXiv:2110.08484 (2021).

[22] J. Xing, J. Liu, J. Wang, L. Sun, X. Chen, X. Gu, Y. Wang, A survey of efficient fine-tuning methods for vision-language models—prompt and adapter, Computers & Graphics 119 (2024) 103885.

[23] M. Chaichuk, S. Gautam, S. Hicks, E. Tutubalina, Prompt to Polyp: Medical Text-Conditioned Image Synthesis with Diffusion Models, arXiv (2025). doi:10.48550/arXiv.2505.05573.

[24] E. Hasanpour Zaryabi, L. Moradi, B. Kalantar, N. Ueda, A. A. Halin, Unboxing the black box of attention mechanisms in remote sensing big data using xai, Remote Sensing 14 (2022) 6254.