Do Medical Vision Encoders Help? PEFT Benchmarking of Vision-Language Models for Gastrointestinal VQA

Matheus Mattioli[1,*], Jurandy Almeida[1]

[1]Intelligent Systems and Data Science Laboratory (LaSID), Federal University of São Carlos (UFSCar), Brazil

Abstract

We present the LaSID–UFSCar participation in the MediaEval Medico 2026 challenge, Subtask 1: Visual Question Answering for Gastrointestinal Imaging (GI-VQA). Starting from Granite Vision 3.2-2B as our primary candidate, we first investigated whether replacing its default SigLIP vision encoder with frozen medical CLIP encoders (MedCLIP, PubMedCLIP) could improve GI-domain alignment. Finding that the default encoder consistently outperformed the medical variants, we broadened the study to a cross-model comparison of five vision-language model families at the ∼2B parameter scale under LoRA, AdaLoRA, and DoRA fine-tuning on a combined medical VQA corpus. Based on those results, we submitted two runs to Kvasir-VQA-x1: Granite Vision 3.2-2B with LoRA and SmolVLM2-2.2B with LoRA. Our best submission—Granite Vision 3.2-2B fine-tuned with LoRA directly on Kvasir-VQA-x1—achieves a public leaderboard BLEU of 0.4928, ROUGE-1 of 0.7182, and METEOR of 0.6968, demonstrating that domain-specific fine-tuning with standard LoRA is highly effective for GI-VQA.

1. Introduction and Related Work

Automated Visual Question Answering (VQA) on endoscopic images can assist clinicians in the rapid interpretation of gastrointestinal (GI) findings. The MediaEval Medico 2026 challenge [1] defines Subtask 1 as generating free-text answers to questions about GI endoscopy images from the Kvasir-VQA-x1 dataset [2], which contains 6,500 original and 65,000 augmented images paired with 159,549 question–answer pairs of varying complexity.

Recent work on medical VQA has explored vision-language pre-training (BLIP [3], BLIP-2 [4]), domain-specific encoders (MedCLIP [5], PubMedCLIP [6]), and instruction-tuned models [7]. Low-Rank Adaptation (LoRA) [8] and its variants, AdaLoRA [9] and DoRA [10], have emerged as the dominant parameter-efficient fine-tuning (PEFT) methods for large vision-language models.

Our candidate model was Granite Vision 3.2-2B (`ibm-granite/granite-vision-3.2-2b`), an IBM enterprise-grade vision-language model (VLM) that pairs a SigLIP vision tower with the Granite 3.1-2B-Instruct language decoder via a LLaVA-style projector, and has demonstrated strong results on document and chart understanding benchmarks. We first investigated whether swapping its generalist SigLIP encoder for a frozen medical CLIP encoder (MedCLIP [5] or PubMedCLIP [6]) could improve medical-domain alignment. Finding that the default SigLIP tower consistently outperformed the medical variants, we broadened the comparison to four additional VLM families to place Granite Vision's baseline performance in context.

Our contributions are: (i) a Granite Vision encoder ablation demonstrating that domain-specific medical CLIP encoders do not improve over the default SigLIP tower under the PEFT regimes studied; (ii) a cross-model comparison of five VLM families at the ∼2B scale on a

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
*Corresponding author.
matheus.mattioli@estudante.ufscar.br (M. Mattioli); jurandy.almeida@ufscar.br (J. Almeida)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

combined medical VQA corpus; and (iii) two competition runs on Kvasir-VQA-x1, with our best submission achieving a public leaderboard BLEU of 0.4928 and ROUGE-1 of 0.7182.

2. Methodology

2.1. Granite Vision and Medical Encoder Adaptation

We hypothesised that replacing Granite Vision's generalist SigLIP encoder with a frozen medical CLIP encoder could improve alignment on GI endoscopy images, and tested two variants: MedCLIP [5] (`ZiyueWang/med-clip`) and PubMedCLIP [6] (`flaviagiammarino/pubmed-clip-vit-base-patch32`). The replacement tower bridges the output dimensionality to the Granite LLaVA-style projector via a trainable linear *DimAdapter* layer ($d_{\text{clip}} \rightarrow d_{\text{SigLIP}}$=1152). Only the *DimAdapter* and the LoRA parameters were updated during training. We also applied LoRA, AdaLoRA, and DoRA on the PubMedCLIP variant as a PEFT ablation. All runs in this study used rank r=16, α=32, dropout 0.05, learning rate 2×10^{-4}, cosine schedule with warmup ratio 0.05, 3 epochs, gradient accumulation steps 8, and bfloat16 precision on a single NVIDIA RTX 5000 Ada (32 GB). Experiments were conducted on a combined medical VQA corpus assembled from the official training splits of SLAKE, VQA-MED, PATH-VQA, and VQA-RAD. The corresponding test splits of the same datasets were merged to form the evaluation set. In the ablation studies, the impact of the medical vision encoders was tested on a randomly sampled 500-question subset of this merged test set, while the remaining experiments for native SigLIP runs and the cross-model comparison used the full test set.

2.2. Cross-Model Comparison

Finding that the default SigLIP encoder outperformed the medical CLIP variants, we expanded the evaluation to four additional VLM families to contextualise Granite Vision's performance:

- **BLIP-2** (`Salesforce/blip2-opt-2.7b`) + LoRA
- **Qwen2-VL-2B** (`Qwen/Qwen2-VL-2B-Instruct`) + LoRA
- **SmolVLM2-2.2B** (`HuggingFaceTB/SmolVLM2-2.2B-Instruct`) + LoRA
- **MedVLM-R1** (`JZPeterPan/MedVLM-R1`), zero-shot only

All models were evaluated on the same combined corpus using identical PEFT hyperparameters as in Section 2.1.

2.3. Competition Submissions

Based on the preliminary results, we submitted two runs to the competition:

- **Run 1**: Granite Vision 3.2-2B + LoRA (r=16, α=32), fine-tuned directly on Kvasir-VQA-x1 (no warm-start), 4 epochs on 4 × RTX 5000 Ada.
- **Run 2**: SmolVLM2-2.2B + LoRA (r=16, α=32), warm-started from the combined-medical adapter, then fine-tuned on Kvasir-VQA-x1 on 1 × RTX 5000 Ada.

All runs applied CLAHE contrast enhancement and online augmentation (rotation $\pm5°$, brightness/contrast factor 0.85–1.15) and used greedy decoding with `max_new_tokens`= 128.

3. Results and Analysis

3.1. Granite Vision Encoder Ablation

Table 1 reports the Granite Vision 3.2-2B encoder ablation on the combined medical VQA test set. Contrary to our initial hypothesis, the default SigLIP tower with LoRA (ROUGE-1 = 0.4468) consistently outperforms both MedCLIP and PubMedCLIP variants. Among medical-encoder configurations, PubMedCLIP + LoRA achieves the best ROUGE-1 of 0.4200, still 2.7 points below the SigLIP baseline. The PEFT ablation on PubMedCLIP shows no clear winner among LoRA, DoRA, and AdaLoRA, indicating that the encoder representation gap is the dominant factor rather than the choice of PEFT method.

Table 1
Granite Vision 3.2-2B encoder ablation on the combined medical VQA test set. Bold: best per column.

PEFT	Vision Enc.	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
LoRA	**default**	**0.0156**	**0.4468**	**0.0192**	**0.4469**	**0.2668**
LoRA	MedCLIP	0.0127	0.3959	0.0147	0.3960	0.2304
LoRA	PubMedCLIP	0.0128	0.4200	0.0181	0.4196	0.2543
DoRA	PubMedCLIP	0.0129	0.3963	0.0165	0.3939	0.2365
AdaLoRA	PubMedCLIP	0.0064	0.4135	0.0114	0.4130	0.2293

3.2. Cross-Model Comparison

Table 2 compares all five VLM families using their default vision encoders. SmolVLM2 + LoRA achieves the highest ROUGE-1 (0.5727) and METEOR (0.3031), closely followed by Qwen2-VL. Granite Vision ranks third among fine-tuned models, ahead of BLIP-2 by a wide margin. MedVLM-R1 in zero-shot configuration produces near-random outputs on this domain, underscoring the importance of fine-tuning even for medically pre-trained models. These results guided our submission strategy: Granite Vision as the original target model and SmolVLM2 as the strongest model from the cross-model comparison.

Table 2
Cross-model comparison on the combined medical VQA test set (default vision encoder; all fine-tuned models use LoRA, r=16, α=32). Bold: best per column.

Model	PEFT	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
MedVLM-R1	zero-shot	0.0002	0.0001	0.0371	0.0370	0.0210
BLIP-2	LoRA	0.0033	0.0781	0.0017	0.0777	0.0428
Granite 3.2-2B	LoRA	0.0156	0.4468	0.0192	0.4469	0.2668
Qwen2-VL-2B	LoRA	0.0619	0.5554	0.0323	0.5549	0.2936
SmolVLM2-2.2B	**LoRA**	**0.0636**	**0.5727**	**0.0351**	**0.5725**	**0.3031**

3.3. Subtask 1 Leaderboard Results

Table 3 reports the public test-set scores for our two competition runs. Granite Vision 3.2-2B with LoRA (Run 1) yields the best scores across all metrics despite ranking lower than SmolVLM2 in the preliminary comparison. We attribute this reversal to two factors: (i) Run 1 was trained with 4 GPUs for 4 epochs directly on Kvasir-VQA-x1, providing longer exposure to the target

domain; and (ii) Granite Vision may generalise better to the longer, more complex answers in Kvasir-VQA-x1 thanks to its 8192-token context window, compared to SmolVLM2's 2048 limit.

Table 3
Public leaderboard scores on Kvasir-VQA-x1 (Subtask 1).

Run	Model	PEFT	Vision Enc.	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
1	**Granite 3.2-2B**	**LoRA**	**default**	**0.4928**	**0.7182**	**0.5393**	**0.6920**	**0.6968**
2	SmolVLM2-2.2B	LoRA	default	0.4215	0.6863	0.4920	0.6569	0.6604

4. Discussion and Conclusion

Our results show that a well-tuned LoRA adapter on a generalist vision-language model can be highly competitive for domain-specific medical VQA. The cross-model comparison revealed that modern 2B-parameter models –SmolVLM2, Qwen2-VL and Granite Vision–, outperform the earlier BLIP-2 generation and zero-shot specialist models on open-ended answer generation.

Why the swapped-encoder Granite Vision underperforms. Replacing the SigLIP visual backbone with a frozen medical CLIP encoder and a linear *DimAdapter* failed to improve performance despite the domain specificity of MedCLIP and PubMedCLIP. We identify two likely causes. First, the *DimAdapter* addresses the representation *dimensionality* mismatch but cannot compensate for the distributional shift between SigLIP patch features and CLIP patch features: the downstream LLaVA-style projector inside Granite Vision was calibrated on SigLIP representations, so the adapted visual tokens remain out-of-distribution for it. Second, and crucially, the Granite language decoder was kept entirely *frozen* throughout these experiments — only the *DimAdapter* and LoRA layers were trained. As a result, even if the visual tokens were perfectly adapted, the decoder had no opportunity to learn the specialized vocabulary and clinical phrasing characteristic of GI VQA. The interplay of these two frozen components likely explains the $\approx$15–17 ROUGE-1 point gap relative to the native-encoder Granite run.

Future work. Several directions could extend this work. First, within the PEFT paradigm, jointly fine-tuning the *DimAdapter*, projector, and language decoder with LoRA — rather than keeping the decoder frozen — is the most direct fix for the swapped-encoder underperformance. Second, the Granite framework could be further improved by replacing the language backbone with a biomedically pre-trained decoder (e.g., BioMedLM, Meditron), creating a fully domain-adapted VLM from the encoder to the output layer. Third, training-free adaptation strategies represent a complementary direction: methods such as test-time augmentation ensembling [11], in-context pseudo-label refinement [12], and vocabulary-free caption retrieval [13] could provide competitive baselines without any gradient-based training, enabling rigorous benchmarking of PEFT against zero-parameter-overhead inference at the domain boundary. Finally, retrieval-augmented generation for rare GI findings and explicit chain-of-thought prompting for multi-part compositional questions are natural extensions within the generative setting.

Acknowledgments

This research was supported by São Paulo Research Foundation - FAPESP (#2023/17577-0, #2025/24680-8) and National Council for Scientific and Technological Development - CNPq

(#315220/2023-6, #420442/2023-5, #444982/2024-8). Experiments were performed on hardware provided by the *LaSID* lab at UFSCar. The authors thank the MediaEval Medico 2026 organisers for the dataset and evaluation infrastructure. Code is available at https://github.com/matheustmattioli/medvqa-mediaeval and the fine-tuned model adapter at https://huggingface.co/mthsmtt/granite-vision-kvasir-vqa.

Declaration on Generative AI

During the preparation of this work the authors used GitHub Copilot to assist with code completion and text editing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Gautam, V. Thambawita, M. Riegler, P. Halvorsen, S. Hicks, Medico 2026: Visual question answering for gastrointestinal imaging, arXiv preprint arXiv:2508.10869 (2025).

[2] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-VQA-x1: A multimodal dataset for medical reasoning and robust MedVQA in gastrointestinal endoscopy, arXiv preprint arXiv:2506.09958 (2025).

[3] J. Li, D. Li, C. Xiong, S. C. H. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: Proceedings of the 39th International Conference on Machine Learning (ICML), 2022.

[4] 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), 2023.

[5] Z. Wang, Z. Wu, D. Agarwal, J. Sun, MedCLIP: Contrastive learning from unpaired medical images and text, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022.

[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, Dubrovnik, Croatia, May 2–6, 2023, 2023.

[7] X. Zhang, C. Wu, 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.

[8] E. J. Hu, Y. Shen, P. Wallis, et al., LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR), 2022.

[9] Q. Zhang, M. Chen, A. Bukharin, et al., AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning, in: International Conference on Learning Representations (ICLR), 2023.

[10] S.-Y. Liu, C.-Y. Wang, H. Yin, et al., DoRA: Weight-decomposed low-rank adaptation, in: Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

[11] M. Farina, G. Franchi, G. Iacca, M. Mancini, E. Ricci, Frustratingly easy test-time adaptation of vision-language models, in: Advances in Neural Information Processing Systems (NeurIPS), volume 37, 2024, pp. 129062–129093.

[12] M. Garosi, M. Farina, A. Conti, M. Mancini, E. Ricci, Large multimodal models as general in-context classifiers, in: Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026.

[13] A. Conti, E. Fini, M. Mancini, P. Rota, Y. Wang, E. Ricci, Vocabulary-free image classification and semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence (2026).