ME-VLIP: A Modular and Efficient Vision-Language Framework for Generalizable Medical Image Parsing

Mai A. Shaaban[1,3,*][0000−0003−1454−6090], Amal Saqib[2][0009−0006−7462−0386], Shahad Hardan[1][0000−0002−4097−1982], Darya Taratynova[1][0009−0005−8344−7709], Tausifa Jan Saleem[2][0000−0002−0827−0043], and Mohammad Yaqub[2][0000−0001−6896−1105]

[1] Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2] Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[3] Department of Mathematics and Computer Science, Faculty of Science, Alexandria University, Alexandria, Egypt
{mai.kassem}@mbzuai.ac.ae

Abstract. Medical image parsing presents a unique challenge due to the diversity of imaging modalities and the wide range of diagnostic tasks required in clinical workflows, including classification, detection, and report generation. Traditional approaches often rely on task-specific models, which limit both scalability and generalization. Recent advances in vision-language models (VLMs) offer promising avenues for unifying these tasks; however, many existing solutions suffer from high computational costs and limited adaptability. In this work, we propose ME-VLIP, a modular and efficient framework built upon InternVL3-8B, fine-tuned using quantized low-rank adaptation and guided by a zero-shot task classification module. Our system demonstrates robust performance across seven tasks, spanning eight imaging modalities. We evaluate our approach on the FLARE 2025-Task 5 benchmark, showing substantial performance gains over the base model, with the following task-specific improvements: classification (0.74 balanced accuracy), multi-label classification (0.57 F1 score), detection (0.82 F1 score), cell counting (251.6 MAE), regression (11.84 MAE), and report generation (0.71 GREEN score). Comparative analysis indicates that our method outperforms other state-of-the-art VLMs, underscoring the effectiveness of parameter-efficient domain adaptation for versatile medical image parsing.

Keywords: Medical image parsing · Parameter-efficient fine-tuning · Vision-language model · Medical visual question answering.

1 Introduction

Medicine inherently involves understanding and reasoning across multiple modalities, notably clinical images and structured or natural-language patient data.

The fusion of image and text enables more nuanced clinical decision-making, particularly in medical visual question answering (MedVQA), where interpretable responses grounded in imaging are essential for accurate diagnosis and treatment guidance [14]. Beyond modality fusion, multi-task learning is crucial in clinical AI: healthcare involves a diverse set of imaging types (e.g. X-ray, ultrasound, CT, microscopy) and diagnostic objectives, from disease classification to lesion detection and automated report generation. A unified multi-task model reduces redundancy and encourages feature sharing across tasks, increasing generalizability across modalities and clinical abnormalities while reducing deployment complexity [3].

Developing a model that is simultaneously 1) multimodal, 2) multi-task, 3) efficient, and 4) generalizable can push the boundaries of AI in medical settings, despite the multiple challenges involved. To this end, FLARE 2025-Task 5[4], a MICCAI 2025 challenge, aims to advance the development of generalist models for multimodal medical image parsing. The FLARE challenge provides a large-scale benchmark with more than 58,000 image-question pairs that cover diverse imaging modalities (e.g., X-ray, ultrasound, microscopy) and multiple tasks, as illustrated in Figure 1, including classification, detection, cell counting, regression, and report generation, enabling a rigorous evaluation of flexibility and scalability in medical AI systems.

Contributions The main contributions are detailed as follows:

- We propose ME-VLIP, a **M**odular and **E**fficient **V**ision-**L**anguage Framework for Generalizable Medical **I**mage **P**arsing.
- We develop a memory- and compute-efficient fine-tuning strategy for InternVL3-8B [20] using Quantized low-rank adaptation (QLoRA) [4]. This strategy achieves domain adaptation under limited hardware while retaining diagnostic performance.
- We introduce a zero-shot task classification (TC) module that dynamically routes image-question pairs to task-specific or multi-task QLoRA adapters at inference time, thereby enhancing task specialization while preserving generalization across modalities.
- We perform a comprehensive evaluation on FLARE 2025-Task 5 benchmark, demonstrating consistent improvements over base models and competitive state-of-the-art VLMs across all tasks and modalities.

Related Work Recent efforts have increasingly focused on vision–language models (VLMs) tailored to clinical domains. These models align visual representations of medical images with language through dedicated fusion modules and pre-training objectives, enabling them to address diverse clinical tasks within a unified multimodal architecture [13]. However, the high resource cost of large models can be prohibitive in real-world clinical environments where latency,

[4] https://www.codabench.org/competitions/7151/

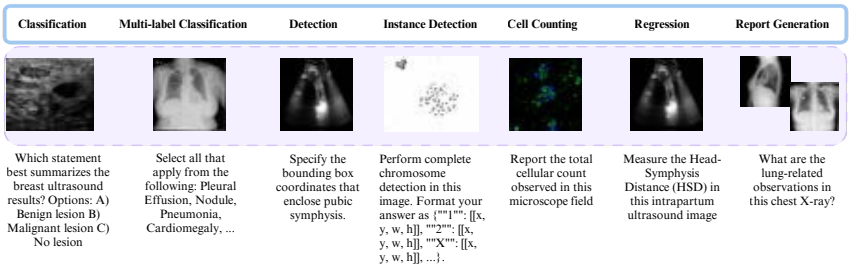


Fig. 1: Overview of the seven clinical tasks in FLARE 2025-Task 5, showing representative input images and associated prompts across different imaging modalities.

compute, and power are constrained. Thus, adopting efficient architectures, such as parameter-efficient adapters or smaller fusion modules, is essential for scalable deployment while preserving diagnostic performance [16]. Both LoRA [6] and QLoRA [4] exemplify the growing trend of parameter-efficient fine-tuning, where large VLMs can be adapted to specialized clinical tasks without retraining the full model. These approaches have been successfully applied in recent works. For example, Gautam et al. [5] fine-tuned Qwen2.5-VL-7B-Instruct using LoRA to jointly perform detection, localization, and counting via instruction-based prompts, simulating clinical reasoning workflows. Their unified approach achieved significant performance gains across all tasks compared to the base model, including improving mean average precision from 0.01 to 0.85 and pointing accuracy from 0.43 to 0.99. These results highlight the benefits of multitask, multimodal fine-tuning for interoperability and generalization in clinical applications. Similarly, UMIT [18] presents a unified VLM tailored to a wide range of medical imaging tasks, including VQA, disease classification, and report generation. UMIT employs a novel two-stage training strategy comprising a feature alignment phase using domain-specific image–text pairs and an instruction fine-tuning phase with medical task templates. This approach enables the model to achieve state-of-the-art performance across benchmarks, including an accuracy of 89. 2% on SLAKE and 95.4% on PathMNIST.

Beyond VQA, other works have focused on adapting foundation models for segmentation and object-level tasks. SAC [10] introduces Segment Any Cell, a framework designed to improve nuclei segmentation with SAM through automated prompt generation and LoRA-based attention adaptation. SAC significantly improves Dice scores on the MoNuSeg dataset, outperforming both fine-tuned SAM [7] and the Medical SAM Adapter. Based on the need for domain-specific adaptation, [15] systematically evaluates large VLM (ChatGPT, Gemini, and LLaVA) alongside SAM for microscopy image analysis. By tuning SAM's parameters for microscopy datasets, they demonstrate substantial improvements in segmentation and counting performance. In particular, the count task R^2 im-

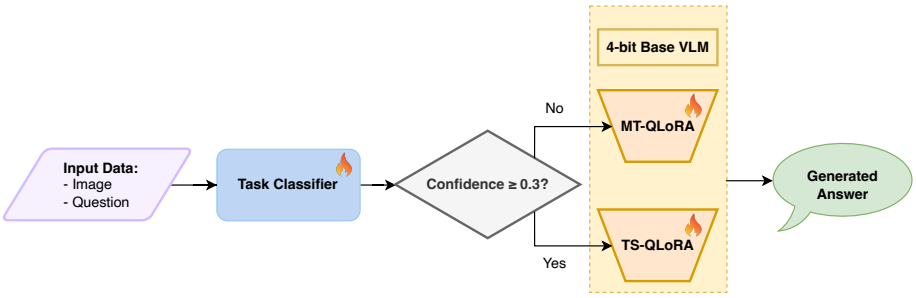


Fig. 2: Overview of ME-VLIP. An input question is first processed by the task classifier to predict its task type. If the classifier's confidence score meets or exceeds a predefined threshold, the corresponding task-specific (TS) QLoRA adapter is loaded and merged with the frozen base model. Otherwise, a generalist multi-task (MT) QLoRA adapter is used. The visual and textual inputs are then processed by the combined model to generate the final prediction.

proves from 0.02 to 0.98 on BBBC005, and the computational segmentation approach of ChatGPT achieves Dice scores between 0.96 and 0.99, outperforming the baseline configuration of SAM.

2 Method

We introduce ME-VLIP, a dynamic, task-aware adaptation mechanism for medical image parsing. Our method leverages efficient QLoRA fine-tuning and a task classification module to specialize model behavior for diverse MedVQA tasks. The system dynamically routes inputs to specialized or generalist adapters based on the predicted task type, optimizing both performance and efficiency, as depicted in Figure 2.

2.1 Efficient Fine-Tuning with QLoRA

We fine-tune InternVL3-8B by adopting QLoRA [4], which enables memory-efficient fine-tuning [20]. QLoRA extends the LoRA approach by combining parameter-efficient fine-tuning [9] with low-bit quantization, reducing memory and compute costs while preserving accuracy.

As in LoRA [6], the pre-trained model weights are frozen, and lightweight trainable adapter layers are inserted into transformer blocks. These adapter weights are the only parameters updated during training, enabling task-specific adaptation without modifying the full model.

QLoRA reduces memory usage by storing frozen model weights in 4-bit quantization (NF4), while performing all computations in bfloat16 (BF16). At runtime, NF4 weights are temporarily dequantized into BF16 for forward and backward passes. Gradients are computed only for the adapter parameters, which

remain in BF16. This dual representation balances compact storage with training stability and efficiency.

Formally, for a linear layer, the QLoRA forward pass is expressed as:

$$\mathbf{Y}^{\text{BF16}} = \mathbf{X}^{\text{BF16}} \cdot \text{doubleDequant}(\mathbf{c}_1^{\text{FP32}}, \mathbf{c}_2^{k\text{-bit}}, \mathbf{W}^{\text{NF4}}) + \mathbf{X}^{\text{BF16}} \cdot \mathbf{L}_1^{\text{BF16}} \cdot \mathbf{L}_2^{\text{BF16}} \quad (1)$$

Where:

- $\mathbf{X}^{\text{BF16}}$ is the input in bfloat16,
- $\mathbf{W}^{\text{NF4}}$ are the 4-bit quantized frozen weights,
- $\mathbf{c}_1^{\text{FP32}}, \mathbf{c}_2^{k\text{-bit}}$ are quantization constants for reconstructing full-precision weights,
- $\mathbf{L}_1^{\text{BF16}}, \mathbf{L}_2^{\text{BF16}}$ are the LoRA adapter matrices.
- k-bit refers to the quantization precision (number of bits used to represent each value). For example, $k = 4$ corresponds to 4-bit NormalFloat (NF4) quantization, which is shown to preserve accuracy while reducing memory footprint [4].

$$\begin{aligned} \text{doubleDequant}\big(\mathbf{c}_1^{\text{FP32}}, \mathbf{c}_2^{k\text{-bit}}, \mathbf{W}^{k\text{-bit}}\big) &= \text{dequant}\Big(\text{dequant}\big(\mathbf{c}_1^{\text{FP32}}, \mathbf{c}_2^{k\text{-bit}}\big), \mathbf{W}^{k\text{-bit}}\Big) \\ &= \mathbf{W}^{\text{BF16}} \quad (2) \end{aligned}$$

This allows accurate reconstruction of full-precision weights for computation while retaining compact low-bit storage. Additionally, QLoRA employs paged optimizers that manage optimizer states in a paged format, dynamically offloading them between GPU and CPU to avoid memory spikes and prevent out-of-memory errors during training.

To support dynamic routing during inference, we develop two types of QLoRA adapters:

- **Task-specific (TS)**: The TS-QLoRA adapters are fine-tuned using only the subset of the MedVQA dataset corresponding to a particular task type (e.g., classification or detection), enabling specialized model behavior.
- **Multi-task (MT)**: The MT-QLoRA adapter are trained on the full dataset, enabling generalist performance across tasks.

2.2 Task Classification Module

We introduce the TC module to dynamically identify the underlying task type associated with each medical question. This module is trained on the MedVQA dataset, which contains diverse question-task pairs. We fine-tune a text encoder, using a prompt-based architecture that embeds both the question and the candidate task labels. Special tokens (`«LABEL»` and `«SEP»`) are added to the tokenizer's vocabulary to structure the input. A typical input sequence is formatted as: `[CLS] «LABEL» {Task_Type} «SEP» {Medical_Question} [SEP]`. The model processes this sequence and employs a scoring function (e.g., a simple dot product or cosine similarity) to compute a compatibility score between the encoded question and each candidate label. This score determines the likelihood that the question belongs to a specific task type.

2.3 Dynamic Inference

The ME-VLIP inference pipeline integrates the QLoRA foundation and the TC module. For a given input (image-question pair), the process is:

1. The medical question is passed to the TC module, which predicts its task type and returns a confidence score.
2. Based on the confidence score:
   - If the score $\geq$ a predefined threshold, the system loads the corresponding TS-QLoRA adapter.
   - Otherwise (indicating uncertainty or a novel task type), the system defaults to the MT-QLoRA adapter.
3. The weights of the selected adapter are merged with the frozen base weights of InternVL3-8B.
4. The full model processes the image-question input and generates the final output prediction.

This design ensures computational efficiency by leveraging lightweight adapters and enables flexible adaptation to both known and unforeseen task types, enhancing the model's robustness and applicability in diverse medical scenarios.

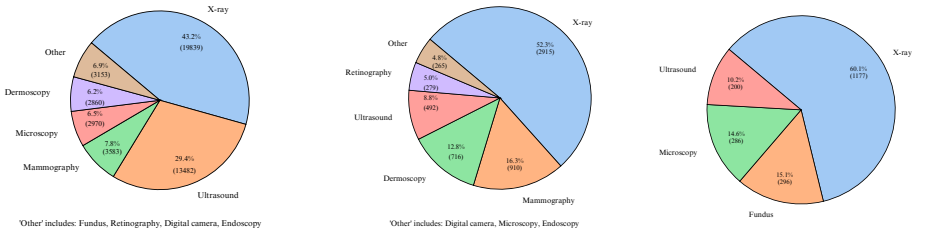
3 Experiments

3.1 Dataset and Evaluation measures

Dataset The FLARE 2025-Task 5 [2] dataset contains eight imaging modalities: clinical, dermatology, endoscopy, mammography, microscopy, retinography, ultrasound, and X-rays. It includes 45,887 questions in the training set, 5,577 in the public validation set, and 1,959 in the hidden validation set (i.e., ground truth answers are not released). The modality distribution across splits is shown in Figure 3. The most represented modality in the dataset is X-ray, followed by ultrasound, mammography, and microscopy.

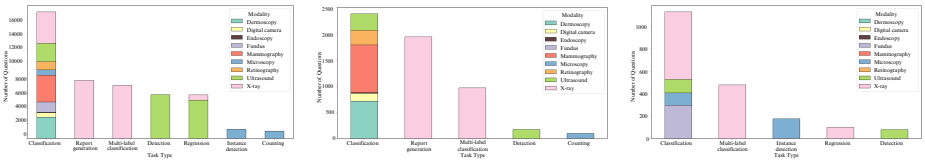
The dataset covers seven task types: classification, cell counting, detection, multi-label classification, regression, and report generation. The modality distribution within each task is illustrated in Figure 4. Notably, classification is the most represented task across all splits and includes the most diverse set of modalities. In contrast, report generation and multi-label classification are limited to X-ray. Detection is restricted to ultrasound, and regression appears primarily with X-ray and a small portion with ultrasound.

Evaluation Each task is assessed using a specific metric depending on its objective. For classification, the evaluation relies on *Balanced Accuracy*, which compensates for the class imbalance by averaging recall between all classes. Multi-label classification is measured using the *Micro-averaged F1 Score*, which combines true positives, false positives, and false negatives in all classes to calculate overall precision, recall, and F1. Detection tasks are evaluated using the



(a) Training set (b) Public validation set (c) Hidden validation set

Fig. 3: Distribution of imaging modalities across the training, public validation, and hidden validation splits.



(a) Training set (b) Public validation set (c) Hidden validation set

Fig. 4: Distribution of imaging modalities per task type across dataset splits. Each bar shows the number of questions per task, stacked by modality.

F1 Score, where predictions are matched to ground truth annotations using an Intersection-over-Union (IoU) threshold of 0.5. Both cell counting and regression tasks use the *Mean Absolute Error (MAE)*, which quantifies the average magnitude of error between predicted and true values. For the report generation task, the evaluation is conducted using *GREEN score*, a domain-specific metric that assesses both the factual correctness and clinical relevance of the generated radiology reports by taking into account lexical overlap, semantic consistency, and clinical accuracy [11].

3.2 Implementation Details

Environment Settings The development environments and requirements are presented in Table 1.

Training Protocols We fine-tuned InternVL3-8B [20] using the LLaMA-Factory framework [19], employing QLoRA (4-bit quantization) with rank of 8 applied to all linear projection layers (q_proj, k_proj, v_proj, o_proj, up_proj, gate_proj, and down_proj). Training followed a Supervised Fine-Tuning (SFT) approach with sequences truncated to 2048 tokens, run for 3 epochs on a single NVIDIA A100-SXM4-40GB GPU (64 CPU cores, 40GB VRAM), using a per-device batch size of 2, gradient accumulation over 4 steps, AdamW optimizer, BF16 precision,

Table 1: Environment and system settings.

Component	Setting
System	Ubuntu 22.04.4 LTS
Programming language	Python 3.10
Dependencies	torch 2.3.1, torchvision 0.18.1, transformers 4.52.dev0
GPU	1x NVIDIA A100-SXM4
VRAM	40GB
CPU	64 cores
Code	https://github.com/BioMedIA-MBZUAI/FLARE2025-Task5-2D-biomedia

Table 2: Training protocols and hyperparameters for QLoRA fine-tuning.

Component	Setting
QLoRA fine-tuning	
Base model	InternVL3 [20]
Number of parameters	8B
Framework	LLaMA-Factory [19]
Method	QLoRA (4-bit quantization)
LoRA rank	8
LoRA target modules	q_proj, k_proj, v_proj, o_proj, up_proj, gate_proj, down_proj
Fine-tuning approach	SFT
Epochs	3
Batch size (per device)	2
Gradient accumulation	4 steps
Effective batch size	8
Sequence length	2048 tokens
Optimizer	AdamW
Precision	BF16
Learning rate schedule	Linear
Initial learning rate	2e-4
Warm-up ratio	3%
Training time	∼60 hours
Inference VRAM	∼9GB
TC configurations	
Model	GLiClass [8]
Encoder	DeBERTa-v3-small
Label model	BGE-small
Epochs	3
Optimizer	AdamW
Loss	Focal Loss ($\alpha = 1$, $\gamma = 1$)
Learning rate	1e-5
Precision	FP16 (mixed)

a linear learning rate schedule with an initial rate of 2e-4, and a 3% warm-up ratio. In our experiments, we applied the same set of training hyperparameters when fine-tuning both the MedGemma-4B [12] and Qwen2.5-VL-7B [1] models.

We implemented our TC module using GLiClass [8] with a DeBERTa-v3-small encoder and BGE-small label model. The model is trained for 3 epochs using AdamW with focal loss ($\alpha = 1, \gamma = 1$), a learning rate of 1e-5, and mixed precision FP16.

The training hyperparameters are presented in Table 2.

4 Results and Discussion

4.1 Results on the Validation Set

The fine-tuned InternVL3-8B model exhibits significant performance gains compared to its base counterpart on MedVQA tasks, as depicted in Figure 5. In classification and detection, where the base model fails entirely (scoring 0.0), the fine-tuned version achieves 0.613 and 0.596, respectively, demonstrating the acquired capability to recognize and localize medical abnormalities. Multi-label classification shows the most dramatic improvement, increasing from 0.026 to 0.510, indicating enhanced ability to handle complex diagnostic labels. Regression and cell counting errors decrease by approximately 40% and 33%, reflecting better precision in quantitative medical assessments. Even report generation, where the base model was moderately competent, fine-tuning yields a 7% improvement.

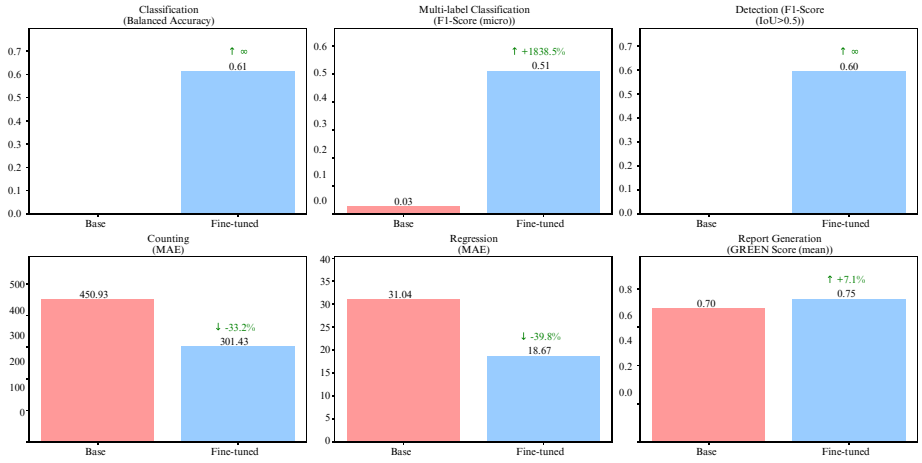


Fig. 5: Performance comparison between the base model and the fine-tuned model. Scores are averaged on public and hidden validation sets.

Table 3: Model performance comparison on validation sets (Public | Hidden scores).

Task & Metric	Qwen2.5-VL-7B	MedGemma-4B	InternVL3-8B	InternVL3-8B (w/ TC)
Classification				
Balanced Accuracy ↑	0.36 | 0.51	0.56 | 0.68	0.52 | 0.71	0.53 | 0.74
Multi-label Classification				
F1 Score ↑	0.36 | 0.53	0.55 | 0.56	0.46 | 0.56	0.46 | 0.57
Detection				
F1 Score ↑	0.64 | 0.71	0.12 | 0.73	0.37 | 0.82	0.26 | 0.82
Instance Detection				
F1 Score ↑	- | 0.00	- | 0.00	- | 0.00	- | 0.00
Cell Counting				
MAE ↓	243.6 | -	265.7 | -	301.4 | -	251.6 | -
Regression				
MAE ↓	- | 15.70	- | 15.35	- | 18.67	- | 11.84
Report Generation				
GREEN Score ↑	0.76 | -	0.74 | -	0.75 | -	0.71 | -

Table 3 compares the performance of three fine-tuned VLMs (Qwen2.5-VL-7B, MedGemma-4B, and InternVL3-8B) across multiple tasks, with InternVL3-8B evaluated both with and without TC. Overall, MedGemma-4B and InternVL3-8B (especially the TC variant) outperform Qwen2.5-VL-7B in most tasks. For classification, InternVL3-8B achieves the highest balanced accuracy (0.71 on hidden validation data), while its TC-enhanced version slightly improves this score to 0.74. In multi-label classification, MedGemma-4B performs consistently well on both public and hidden validation sets, whereas InternVL3-8B excels only on the hidden set. This discrepancy likely stems from MedGemma's pre-training on large-scale medical datasets, contrasting with InternVL's generic pre-training, though the performance gains are marginal (F1 scores of 0.56–0.57). Detection tasks reveal a stark contrast: Qwen2.5-VL-7B performs well on public data (F1=0.64), but InternVL3-8B with TC dominates on hidden data (F1=0.82), suggesting superior generalization. Notably, all models fail completely in instance detection (F1=0.00), highlighting a critical limitation in fine-grained localization. In regression tasks, InternVL3-8B with TC achieves the lowest MAE (11.84), significantly outperforming the others and underscoring the benefits of task-specific adaptation. For cell counting, Qwen2.5-VL-7B performs best (MAE=243.6), though the high errors across all models indicate this task remains challenging. Report generation, measured by GREEN [11] score, shows minimal differences, with Qwen2.5-VL-7B slightly ahead (0.76).

Ablation Study An ablation study comparing the fine-tuned InternVL3-8B with and without TC in Table 3 demonstrates that task-specific adaptations often yield superior performance. These improvements underscore the necessity of domain adaptation for VLMs in medical applications. While the pre-trained model possesses general capabilities, its lack of medical-specific knowledge limits diagnostic utility. Hence, properly adapted models could serve as valuable assis-

tants in medical imaging analysis, though their effectiveness depends critically on the quality and breadth of training data.

4.2 Results on the Final Testing Set

This is a placeholder. We will present the testing results during MICCAI.

4.3 Limitation and Future Work

The persistent challenges in instance detection and cell counting highlight areas where current methods and our proposed method still fall short, pointing to a need for improved architectural designs or training strategies for fine-grained visual understanding and numerical reasoning.

Future work will focus on optimizing the fine-tuning protocol for these challenging tasks, exploring advanced adapter architectures, and validating the model's robustness across an even broader spectrum of medical modalities and clinical scenarios. This work serves as a step towards scalable and versatile AI assistants that can be effectively integrated into heterogeneous clinical workflows.

5 Conclusion

This study presented a unified, efficient vision-language framework for multimodal medical image parsing, addressing the challenges of task diversity, modality integration, and scalability. By leveraging QLoRA for memory-efficient fine-tuning and a zero-shot task classifier for dynamic adapter routing, our system adapts flexibly to varied clinical tasks while maintaining computational feasibility.

Experimental results on the FLARE 2025-Task 5 benchmark demonstrate significant performance gains across classification, detection, cell counting, regression, and report generation tasks. Notably, our InternVL3-8B-based model with task-aware QLoRA adapters achieves state-of-the-art results, underscoring the value of modular adaptation strategies.

Acknowledgements The authors of this paper declare that the proposed solution is fully automatic without any manual intervention. We thank all data owners and contributors for making the data publicly available and CodaLab [17] for hosting the challenge platform.

Disclosure of Interests

The authors declare no competing interests.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 9
2. Codabench: Miccai flare 2025 task 5: Multimodal model for medical image parsing (2025), https://www.codabench.org/competitions/7151/ 6
3. Demirhan, H., Zadrozny, W.: Survey of multimodal medical question answering. BioMedInformatics **4**(1), 50–74 (2024), https://www.mdpi.com/2673-7426/4/1/4 2
4. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems **36**, 10088–10115 (2023) 2, 3, 4, 5
5. Gautam, S., Riegler, M.A., Halvorsen, P.: Point, detect, count: Multi-task medical image understanding with instruction-tuned vision-language models (2025), https://arxiv.org/abs/2505.16647 3
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022) 3, 4
7. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023) 3
8. Knowledgator: Gliclass: Generalist and lightweight model for sequence classification (2024), https://github.com/Knowledgator/GLiClass 8, 9
9. Liu, Y., Ma, Y., Chen, S., Ding, Z., He, B., Han, Z., Tresp, V.: Perft: Parameter-efficient routed fine-tuning for mixture-of-expert model. arXiv preprint arXiv:2411.08212 (2024) 4
10. Na, S., Guo, Y., Jiang, F., Ma, H., Huang, J.: Segment any cell: A sam-based auto-prompting fine-tuning framework for nuclei segmentation (2024), https://arxiv.org/abs/2401.13220 3
11. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Md, A.E.M., Moseley, M., Langlotz, C., Chaudhari, A.S., Delbrouck, J.B.: Green: Generative radiology report evaluation and error notation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. p. 374–390. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.21, http://dx.doi.org/10.18653/v1/2024.findings-emnlp.21 7, 10
12. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025) 9
13. Shaaban, M.A., Saleem, T.J., Papineni, V.R., Yaqub, M.: Motor: Multimodal optimal transport via grounded retrieval in medical visual question answering (2025), https://arxiv.org/abs/2506.22900 2
14. Tu, T., Azizi, S., et al.: Towards generalist biomedical AI (2023), https://arxiv.org/abs/2307.14334 2
15. Verma, P., Van, M.H., Wu, X.: Beyond human vision: The role of large vision language models in microscope image analysis (2024), https://arxiv.org/abs/2405.00876 3

16. Wang, T., Zhou, W., Zeng, Y., Zhang, X.: Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning (2022), https://arxiv.org/abs/2210.07795 3
17. Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 11
18. Yu, H., Yi, S., Niu, K., Zhuo, M., Li, B.: Umit: Unifying medical imaging tasks via vision-language models (2025), https://arxiv.org/abs/2503.15892 3
19. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., Ma, Y.: Llamafactory: Unified efficient fine-tuning of 100+ language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Association for Computational Linguistics, Bangkok, Thailand (2024), http://arxiv.org/abs/2403.13372 7, 8
20. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) 2, 4, 7, 8

Table 4: Checklist Table. Please fill out this checklist table in the answer column.

Requirements	Answer
A meaningful title	Yes
The number of authors (≤6)	6
Author affiliations and ORCID	Yes
Corresponding author email is presented	Yes
Validation scores are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	Figure number: 2
Pre-processing	N/A
Strategies to use the partial label	N/A
Strategies to use the unlabeled images.	N/A
Strategies to improve model inference	Page number: 4-6
Post-processing	N/A
The dataset and evaluation metric section are presented	Page number: 6
Environment setting table is provided	Table number: 1
Training protocol table is provided	Table number: 2
Ablation study	Page number: 10
Efficiency evaluation results are provided	Table number: 3
Visualized segmentation example is provided	N/A
Limitation and future work are presented	Yes
Reference format is consistent.	Yes