FeLoRA16-SPP: Parameter-Efficient Fine-Tuning of 3D Multimodal LLMs for Radiology Report Generation and Visual Question Answering

Malaikah Javed[1], Areeb Ahmad Chaudhry[1], Nazia Perwaiz[1], Usama Athar[1], Ameen Salahudeen[2], and Muhammad Moazam Fraz[1]

[1] School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan
[2] University of Illinois, Chicago, USA
nazia.perwaiz@seecs.edu.pk

Abstract. Multimodal Large Language Models (MLLMs) are emerging as powerful tools for automating radiology report generation (RRG) and visual question answering (VQA) from 3D CT scans. In this work, we present FeLoRA16-SPP, a lightweight adaptation of the M3D-LaMed baseline that combines LoRA-based para-meter-efficient fine-tuning with a spatial pooling projector, while freezing the vision encoder for efficiency. We evaluate FeLoRA16-SPP on the FLARE25 MICCAI challenge using the GREEN score, the official metric for organ-level report completeness. Our method improves performance on the GREEN score by up to 12% compared to PHI3 and by 4% compared to Med3DVLM, achieving an average score of 0.431 across 18 organ systems. FeLoRA16-SPP delivers top results for 9 organs including the respiratory tract (0.7901), kidneys (0.3386), biliary system (0.6344), pancreas (0.6173), and lymphatic system (0.6600). These results demonstrate that parameter-efficient adaptations of 3D MLLMs can provide clinically meaningful improvements in structured radiology report generation without requiring full-scale retraining.

Keywords: PEFT · GREEN · Radiology report generation · VQA.

1 Introduction

Foundation models are transforming medical image analysis by enabling downstream adaptability with minimal supervision [2] [9]. Recent work has begun to explore 3D vision-language models for medical images. The M3D framework introduced a large 3D image–text dataset (M3D-Data) and the M3D-LaMed model for tasks including 3D image-to-report generation and VQA [3]. Following this, Med3DVLM proposed an efficient 3D vision encoder and advanced multimodal projector, achieving state-of-the-art results in 3D report generation and VQA [11]. More recently, a vision–language foundation model tailored for 3D medical imaging, further advancing the design space for volumetric vision–language systems was proposed [10]. These works demonstrate that

transformer-based MLLMs can effectively process volumetric CT and MRI data by flattening 3D features into a language model input. For example, a systematic study of the 3D MLLM design space showed that model size and masking strategies influence report generation accuracy, and that combining the raw CT volume with an organ segmentation mask can further improve performance [2].

Multimodal models have also been evaluated on radiology tasks specifically. Benchmarks like AMOS-MM and CT2Rep emphasize report generation accuracy, while other works focus on visual question answering for 3D scans [12]. Surveys of medical VLMs highlight that recent models integrate advances from general vision–language research (e.g., CLIP-style pretraining, MLP-Mixer projectors, contrastive or generative objectives) into medical data. Importantly, these reviews note that RRG and RVQA remain challenging due to limited paired data and the complexity of medical language [12] [4]. Our work builds on these efforts by using the M3D baseline as a starting point and tailoring it for the FLARE5 requirements (i.e., report generation and VQA)

The FLARE 2025 Task 5 (Multimodal Model for 3D Medical Image Parsing) requires a single model to handle both report generation and visual question answering (VQA) from 3D CT scans. This reflects a growing need for generalist AI systems in radiology, where images and associated text (reports, notes) must be interpreted together. A recent review of vision–language foundation models for 3D medical imaging highlights the rapid evolution of architectures, datasets, and evaluation protocols in this emerging field [10]. In particular, radiology report generation (RRG) and visual question answering (VQA) from images are two closely related tasks with direct clinical impact [4]. Generating accurate reports can reduce radiologist workload and support under-resourced settings, while VQA allows interactive querying of 3D scans to clarify findings (e.g., "Is there a tumor in the pancreas?"). These tasks leverage recent advances in multimodal large language models (MLLMs) that combine image features with powerful LLMs [11]. The FLARE5 setup combines these modalities and tasks, demanding models that can seamlessly integrate visual and textual information from 3D CTs to answer clinical questions.

The key motivation of this work is to leverage the strong pretraining of 3D MLLMs (e.g., M3D-LaMed) while fine-tuning them for the specific reporting and question-answering tasks of FLARE Task 5. We propose FeLoRA16-SPP, which retains the original 3D MLLM architecture [3] [2] but adapts it with tuned hyperparameters and input preprocessing. Our contributions include a parameter-efficient fine-tuning strategy for 3D medical LLMs that improves GREEN score [7] on report generation for CT scans. Results on the FLARE5 validation sets show that FeLoRA16-SPP outperforms the provided Med3DVLM baseline in GREEN score across most anatomical systems, achieving top scores on biliary tract, pancreas, kidneys, lymphatic system, musculoskeletal system, and respiratory tract. Overall, this work shows that relatively light-touch fine-tuning of a strong 3D MLLM can yield significant gains in clinically relevant subtasks, even if global answer diversity is somewhat reduced.

2 Method

We adopt the LLaVA-style MLLM framework for its simple, effective, and widely used modular design—a visual encoder, a projector, and a language model. In this work, we fine-tuned M3D-LaMed-Phi-3-4B, which is itself an MLLM composed of a ViT-based 3D visual encoder, a projector, and a Phi-3 language model, and was previously pretrained on large-scale medical imaging data. The model's projected visual embeddings are concatenated with the query prompt and supplied to the language model to generate radiology reports. Architecture illustration is provided in Figure 1.

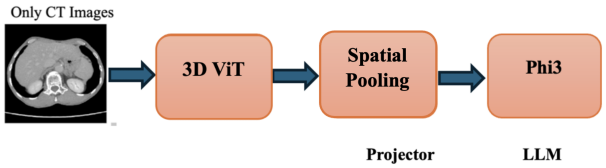

Fig. 1. Overall architecture of the M3D-LaMed-Phi-3-4B model. CT volumes are encoded by a 3D ViT, projected via token pooling, and passed to the Phi-3 language model for report generation.

2.1 Preprocessing

The raw 3D CT volumes were standardized to the model's input format. Intensities were clipped to clinically relevant ranges (-160–240 HU for AMOS/abdominal, -1350–150 HU for others), rescaled to $[0, 1]$, and resampled to $32\times256\times256$ with a single channel. The processed volumes were stored as binary arrays with an updated manifest to ensure reproducible inputs for fine-tuning. Figure 2 shows the preprocessing workflow.

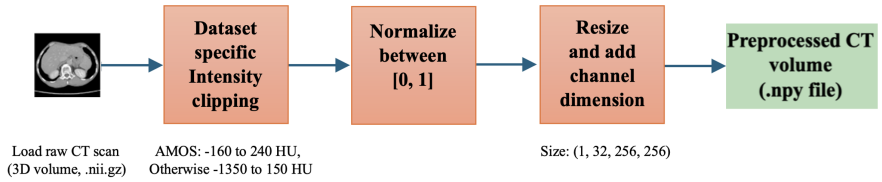

Fig. 2. Preprocessing pipeline. Raw CT volumes in NIfTI format are loaded, intensity values are clipped within dataset-specific Hounsfield Unit ranges, normalized to [0,1], resampled to a fixed resolution ($32\times256\times256$), and stored as standardized .npy files.

2.2 Proposed Method

The overall pipeline of the proposed method is illustrated in Figure 3. We fine-tune the M3D-LaMed-Phi-3-4B model, designed for multimodal medical understanding tasks. A 3D Vision Transformer (ViT) encoder processes the CT volumes, which are preprocessed to a standardized resolution of 32 × 256 × 256 and partitioned into 4 × 16 × 16 voxel patches embedded as tokens. The vision encoder is kept frozen to leverage pretraining on large-scale medical CT datasets.

To bridge the encoder and language model, we experiment with two projector designs. The Spatial Pooling Perceiver (SPP) reduces token count while preserving 3D structure, and the HILT projector uses stacked cross-attention blocks to enrich coarse representations with fine-grained spatial details. Both projectors output refined features via an MLP. The SPP-based model is FeLoRA16-SPP, and the HILT-based model is FeLoRA16-HILT. The fused representations are input into the Phi-3 language model, fine-tuned with LoRA adapters, which generates structured radiology findings, impressions, and answers clinical VQA queries.

Overall, this design balances efficiency and adaptability: by freezing the vision encoder and only fine-tuning the multimodal projector and LLM, less than 8% of the total parameters (321M of 4.07B) are trainable, significantly reducing computational cost while still enabling effective cross-modal alignment and medical report generation.

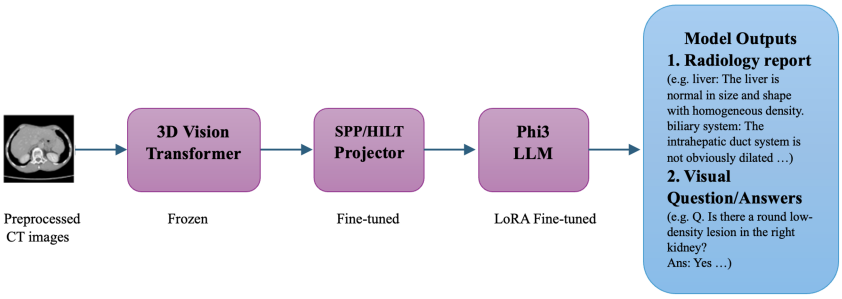

Fig. 3. Proposed model. CT images are encoded by a frozen 3D Vision Transformer, projected into the language space by a projector, and processed by a LoRA-tuned Phi-3 LLM to generate radiology reports and answer clinical questions.

3 Experiments

3.1 Dataset and evaluation measures

The training is conducted using the MICCAI FLARE 2025 Task 5 3D dataset. In total, the training dataset includes 3,290 CT volumes and over 30,000 ques-

tions, supporting both radiology report generation and visual question answering (VQA). The validation set contains 681 samples. Annotations are stored in structured JSON format containing case_id, findings, impression, and both global and local VQA. The dataset combines contributions from CT-RATE (2,000 scans, 19,276 questions) and AMOS-MM (1,236 scans, 12,287 questions), covering chest and abdominal regions with multi-organ details. There are a total of 18 organs covered in the dataset. The official metric of the challenge which is GREEN score [7] is used to evaluate the outputs of report generation and accuracy is used to evaluate the local and global VQA. We also compute text similarity metrics such as BLEU [8], METEOR [5]and ROUGE [6] for report generation.

3.2 Implementation details

The input scan was resized to a volume size of $32 \times 256 \times 256$, ensuring consistent spatial representation across the datasets. For the SPP projector, the pooling type used is spatial pooling, with a pooling size of 2 to reduce the number of tokens while preserving spatial structure. The projector layer type is set to MLP, consisting of 2 layers, which refine the pooled features before passing them to the language model. For the HILT projector, we used 2 stacked cross-attention transformer blocks with 12 attention heads each, followed by a 3-layer MLP to refine low-resolution features using high-resolution context. A prompt-based training strategy was adopted, where a simple template instruction guided the language model to generate descriptive findings and answer clinical questions. Prompts used in training are mentioned in Table 1.

Table 1. Prompts for report generation and VQA tasks.

Task	**Prompt**
Report Generation	You are an AI assistant trained to act as a thoracoabdominal radiologist. Please describe in detail the findings in this thoracoabdominal CT scan. Include all relevant anatomical structures visible in the scan — both thoracic (e.g., lungs, heart, mediastinum) and abdominal (e.g., liver, pancreas, kidneys, etc.).
VQA	You are an AI assistant acting as a radiologist tasked with answering a multiple-choice question based on a CT scan.

Environment settings All model training and inference were implemented using the Hugging Face Transformers framework. The development environments and requirements are presented in Table 2.

Table 2. Development environments and requirements.

System	Ubuntu 22.04 LTS
RAM	48 GB
GPU (number and type)	1 × NVIDIA A40, 48 GB VRAM
CUDA version	11.8
Programming language	Python 3.10.14
Deep learning framework	torch 2.2.1, torchvision 0.17.1

Training protocols The vision encoder of the MLLM was kept frozen, while the projector was finetuned and the LLM (Phi-3) was fine-tuned using LoRA adapters (rank 16). This selective tuning strategy allowed efficient adaptation of cross-modal alignment and language reasoning while significantly reducing the number of trainable parameters. The model was trained for 6 epochs with a batch size of 4, gradient accumulation of 2, which makes effective batch size 8, and a learning rate of 5×10^{-5}, scheduled using a cosine decay strategy with a warmup ratio of 0.03. To optimize memory usage and improve throughput, bfloat16 precision and gradient checkpointing were employed. A summary of the training hyperparameters is presented in Table 3.

Table 3. Training protocols.

Batch size	4
Gradient accumulation	2
Patch size	$4 \times 16 \times 16$
Total epochs	6
Optimizer	AdamW (default parameters)
Initial learning rate (lr)	5×10^{-5}
Lr decay schedule	cosine
Training time	10 hours
Loss function	CrossEntropyLoss
Number of model parameters	4.07B

4 Results and discussion

The quantitative results of radiology report generation are presented in Table 4. FeLoRA16-SPP, demonstrates consistently strong performance across several organ systems. For instance, it achieved high GREEN scores for the Respiratory Tract (0.7901), Esophagus (0.6816), Biliary System (0.6344), Lymphatic System (0.6600), Pancreas (0.6173), Musculoskeletal System (0.5911), Abdominal Cavity and Peritoneum (0.4721), Kidneys (0.3386), and Lungs & Pleura (0.3230). These organs are generally characterized by well-defined anatomical boundaries and consistent appearance in CT scans, which allowed the model to effectively align image features with report text.

Performance was notably low for Blood Vessels (0.0472), Gastrointestinal Tract (0.0790), Breast Tissue (0.0), and Diaphragm (0.0). In particular, breast tissue and diaphragm were largely not captured, reflecting their extreme underrepresentation in the training set (95 and 20 samples out of 3,290, respectively). The complexity and small size of these structures, combined with their scarcity, made them especially challenging to capture. Blood vessels and gastrointestinal structures, though better represented, are highly variable and intertwined with neighboring tissues, leading to reduced sensitivity. Additionally, thin structures like the diaphragm are easily lost during patch-based downsampling, and freezing the vision encoder further limited adaptation to fine-grained spatial features, exacerbating the poor detection of these regions.

FeLoRA16-SPP consistently outperformed the baselines Med3DVLM and PHI3 on multiple organ systems. An alternative configuration, FeLoRA16-HILT, which uses a different projector, showed competitive results in certain cases (e.g., Spleen and endocrine system), but overall FeLoRA16-SPP yielded superior average GREEN score across all regions as presented in Table 4. The SPP projector outperformed HILT for these challenging regions because it preserves the 3D spatial structure, maintaining coarse-to-fine context even for scarce or thin structures. In contrast, HILT's cross-attention sometimes missed these regions when low-resolution queries failed to capture high-resolution details.

The VQA results are shown in Table 5. FeLoRA16-SPP achieved higher accuracy on local questions (0.4054) than global questions (0.0954), reflecting task characteristics: local questions had few well-defined answer choices (e.g., "Yes/No"), constraining the output space and facilitating model performance. Global questions involved many possible conditions, increasing difficulty and reducing accuracy. Compared with baselines, FeLoRA16-SPP performed competitively on local questions but lagged behind three other models, and exhibited lower accuracy on global questions, where Med3DVLM achieved superior results.

Based on the NLP metrics reported in Table 6 for report generation, FeLoRA16-SPP performs better than FeLoRA16-HILT across all metrics. It has higher ROUGE (0.4899 vs 0.4560), METEOR (0.3549 vs 0.3126), and BLEU (0.1598 vs 0.1047), indicating that its generated reports are closer to the ground truth in terms of sequence overlap, semantic similarity, and n-gram precision. This suggests that FeLoRA16-SPP produces more accurate, fluent, and clinically relevant text than FeLoRA16-HILT.

4.1 Quantitative results on validation set

GREEN score comparison is reported for FeLoRA16-SPP, FeLoRA16-HILT and the provided baselines of Med3DVLM and PHI3.

Table 4. Results of radiology report generation (GREEN score).

Organ/System	FeLoRA16-SPP	FeLoRA16-HILT	Med3DVLM	PHI3
Liver	0.2740	0.2278	**0.2775**	0.2408
Biliary System	**0.6344**	0.5986	0.5122	0.4842
Spleen	0.7024	**0.7508**	0.6346	0.5691
Pancreas	**0.6173**	0.58	0.5835	0.5351
Kidneys	**0.3386**	0.2869	0.2470	0.2243
Endocrine System	0.4094	**0.4294**	0.2804	0.2296
Lymphatic System	**0.66**	0.644	0.5643	0.4946
Gastrointestinal Tract	0.079	0.073	**0.1199**	0.0761
Abdominal Cavity & Peritoneum	**0.4721**	0.3462	0.3962	0.3148
Blood Vessels	0.0472	0	**0.1261**	0.1016
Musculoskeletal System	**0.5911**	0.515	0.5309	0.4255
Lungs & Pleura	**0.3230**	0.3038	0.2729	0.2142
Respiratory Tract	**0.7901**	0.7183	0.7794	0.7049
Heart	0.5462	0.3544	0.6545	**0.6663**
Mediastinum	0.6016	0.6025	**0.6117**	0.5192
Esophagus	**0.6816**	0.567	0.6408	0.6626
Breast Tissue	0	0	0	0
Diaphragm	0	0	0	0
Average	**0.431**	0.389	0.402	0.359

Table 5. Comparison results of visual question answering.

Metric	FeLoRA16-SPP	FeLoRA16-HILT	Med3DVLM	PHI3
Global Accuracy	0.0954	0.0859	**0.2358**	0.1799
Local Accuracy	0.4054	0.4441	0.4180	**0.5691**

Table 6. Comparison results of NLP metrics on report generation.

Metric(Avg.)	FeLoRA16-SPP	FeLoRA16-HILT
ROUGE	**0.4899**	0.4560
METEOR	**0.3549**	0.3126
BLEU	**0.1598**	0.1047

4.2 Qualitative results on validation set

Table 7 shows comparison of generated report and the ground truth. The generated report has an average green score among all organs of 0.925. The model

identified every organ that was in the ground truth and nearly correctly capturing their associated findings. While Table 8 presents a generated report with an average green score of 0.06 across all organs. The model exhibited hallucination by including a non-existent organ and produced inaccurate descriptions for the correctly identified organ.

Table 7. The model captured all regions in its report, yielding high GREEN score.

Generated	Ground truth
lungs and pleura: There are subcentimeter nonspecific nodules in both lungs. No mass or infiltrative lesion is identified in either lung. No pleural effusion is present. **respiratory tract**: The trachea and both main bronchi are patent. No occlusive pathology is identified in the trachea or main bronchi. **heart**: The heart is normal in size and contour. No pericardial effusion is present. **mediastinum**: Mediastinal structures cannot be optimally evaluated due to the absence of intravenous contrast. The widths of the mediastinal great vessels are normal. **esophagus**: No abnormal wall thickening of the esophagus is observed within the imaged sections. **lymphatic system**: No pathologically enlarged lymph nodes are identified in the mediastinum or hilar regions. No enlarged lymph nodes of pathological size are detected. **musculoskeletal system**: Thoracic vertebral body heights, alignment, and densities are normal. Intervertebral disc spaces are preserved. Neural foramina are patent. No lytic or destructive bone lesions are identified within the imaged sections. **abdominal cavity and peritoneum**: No free fluid collection is identified in the upper abdomen on the imaged sections. No mass with distinct borders is observed in the upper abdominal organs within the limits of the non-contrast CT.	**lungs and pleura**: There is plaque-like linear calcification in the pleura of the apical segment of the right upper lobe adjacent to the mediastinum, with adjacent parenchymal scarring. No evidence of pneumothorax is seen in either lung. No active infiltration or mass lesion is identified in either lung. Both lungs are well ventilated. No pleural effusion is present. **respiratory tract**: The trachea and both main bronchi are patent without evidence of occlusive pathology. **heart**: Cardiac contour and size are within normal limits as far as can be evaluated without intravenous contrast. No pericardial effusion is detected. **mediastinum**: Mediastinal vascular structures are normal in caliber as far as can be evaluated without intravenous contrast. **esophagus**: No abnormal wall thickening is observed in the thoracic esophagus. **lymphatic system**: No lymph nodes of pathological size or appearance are identified in the mediastinum or axillary regions. **musculoskeletal system**: No lytic or destructive bone lesions are seen. Vertebral body heights are preserved. **abdominal cavity and peritoneum**: No abnormality is detected in the upper abdominal sections within the field of view on non-contrast CT.

Table 8. The model hallucinated a region and produced inaccurate report for the identified organ, resulting in a low GREEN score.

Generated	Ground truth
kidneys: Both kidneys are normal in position, shape, and size with smooth contours. The perirenal fat space is clear. The parenchymal density is uniform without focal density abnormalities. The left renal pelvis, calyces, and ureter are dilated with hydronephrosis. The right renal pelvis, calyces, and ureter are not dilated. The right renal hilum is located anteriorly. **lymphatic system**: No enlarged lymph nodes are seen in the retroperitoneum.	**kidneys**: The left kidney is enlarged. Dilation of the right renal pelvis with fluid accumulation. Thinning of the renal parenchyma. Multiple spotted high-density foci in the collecting system, largest measuring approximately $11\,\mathrm{mm}\sqrt{65}\,\mathrm{mm}$. Multiple cyst-like low-density foci in the left kidney, largest measuring approximately 16mm. Multiple cyst-like low-density foci in the renal parenchyma, largest measuring approximately 15mm. No obvious abnormalities in the position or morphology of the right kidney. The perirenal fat space is clear. No dilation of the renal pelvis or calyces.

Table 9 and Table 10 show qualitative results of global and local visual question answers.

Table 9. Qualitative results of the model on global visual question answers. The model successfully predicted several answers accurately, although some predictions were missing.

Global Question	**Choices (Options Given)**	**Ground Truth Answer**	**Generated Answer**
Which of the following conditions are present in this scan? (Select all that apply, leave empty if none)	Arterial wall calcification, Coronary artery wall calcification, Hiatal hernia, Lymphadenopathy, Emphysema, Atelectasis, Lung nodule, Lung opacity, Pulmonary fibrotic sequela, Pleural effusion, Mosaic attenuation pattern, Peribronchial thickening, Consolidation, Bronchiectasis, Interlobular septal thickening	Arterial wall calcification, Hiatal hernia, Emphysema, Lung nodule	Emphysema, Lung nodule

Table 10. Qualitative results of the model on local visual question answers. The model accurately addressed most root questions and demonstrated strong performance on follow-up questions.

Local Question(s)	**Ground Truth Answer**	**Generated Answer**	**Match**
1. Are there two lesions of cystic density in the right kidney? 2. How would you describe the appearance of these right kidney lesions? Choices: [Solid; Cystic; Calcified]	Yes Cystic	Yes Cystic	Root question correct, Follow-up correct
1. Is there diffuse density loss in the liver? 2. What is the most likely clinical implication of the diffuse density loss in the liver? Choices: [Hepatosteatosis; Cirrhosis; Hepatic congestion]	No Hepatosteatosis	No None of the above	Root question correct, Follow-up incorrect (missed implication)
1. Are there parenchymal findings in the lungs during the late recovery period of Covid pneumonia? 2. Which lung regions show these parenchymal findings? Choices: [Only upper lobes; Only lower lobes; Posterior segment of the upper lobes, right middle lobe, and lower lobes of both lungs]	No Posterior segment of the upper lobes, right middle lobe, and lower lobes of both lungs	Yes Posterior segment of the upper lobes, right middle lobe, and lower lobes of both lungs	Root question incorrect (opposite yes/no), Follow-up correct

4.3 Limitation and future work

Our approach has several limitations. First, freezing the vision encoder, while efficient, likely limited the model's ability to capture fine-grained details, leading to weaker performance in detecting thin or subtle structures such as blood vessels and the diaphragm. Second, we did not perform ablation studies (e.g., varying LoRA rank, frozen vs. unfrozen encoder layers), which restricts our ability to attribute gains to specific design choices. Third, the evaluation metric itself introduces bias: the GREEN score emphasises coverage across major organ systems but may obscure failures in rare or clinically critical findings. Fourth, our model is restricted to CT scans of the abdomen and thorax, limiting generalizability;

extending to other imaging modalities (e.g., MRI, X-ray) will require modality-specific pretraining or domain adaptation. Finally, we relied on the original M3D backbone without exploring newer vision-language designs (e.g., DC-Former [1]) or larger LLMs (e.g., Phi-3), which may further improve representation learning. Future work could address these gaps by conducting systematic ablations of fine-tuning strategies, unfreezing parts of the vision encoder, and incorporating multimodal training with more diverse CT datasets. Designing explicit multi-task heads for detection and counting tasks, along with context-aware mechanisms such as cross-slice attention, may also improve global reasoning. Moreover, supplementing benchmark-based evaluation with expert radiologist assessment will help identify clinically meaningful failure modes beyond what automatic metrics capture.

5 Conclusion

We have presented FeLoRA16-SPP, a fine-tuned adaptation of the M3D 3D-Medical LLM for radiology report generation from 3D CT scans. Our method retains the original M3D backbone while introducing task-specific fine-tuning strategies. Although FeLoRA16-SPP does not surpass Med3DVLM and PHI3 in global or local VQA accuracy, it achieves consistently higher GREEN scores across multiple anatomical systems, including spleen, biliary tract, pancreas, kidneys, lymphatic system, musculoskeletal system, and respiratory tract. These results highlight that our fine-tuning strategy enhances system-level consistency and organ-specific precision in report generation, even when overall answer diversity remains limited. In summary, FeLoRA16-SPP demonstrates the value of lightweight fine-tuning in improving clinically relevant evaluation metrics, particularly the GREEN score, for 3D medical image understanding.

Acknowledgements We thank the FLARE challenge organizers for providing the multimodal dataset and evaluation framework. We also acknowledge the authors of the M3D baseline [2] and Med3DVLM [11] for releasing code and models that enabled our fine-tuning experiments. This research was supported in part by SEECS NUST(National University of Sciences and Technology). Computational resources were provided by MachVIS Lab, SEECS.

Disclosure of Interests

The authors declare no competing financial interests or personal relationships that could influence the work reported in this paper.

References

1. Ates, G.C., Xin, Y., Gong, K., Shao, W.: Dcformer: Efficient 3d vision–language modeling with decomposed convolutions. arXiv preprint arXiv:2502.05091 (2025) 12

2. Baharoon, M., Ma, J., Fang, C., Toma, A., Wang, B.: Exploring the design space of 3d mllms for ct report generation. arXiv preprint arXiv:2506.21535 (2025) 1, 2, 12
3. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024) 1, 2
4. Hartsock, I., Rasool, G.: Vision–language models for medical report generation and visual question answering: a review. Frontiers in Artificial Intelligence **7**, 1430984 (2024) 2
5. Lavie, A., Agarwal, A.: Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation (StatMT '07). pp. 228–231. Association for Computational Linguistics, USA (2007) 5
6. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004) 5
7. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Md, A., Moseley, M., Langlotz, C., Chaudhari, A., et al.: Green: Generative radiology report evaluation and error notation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 374–390 (2024) 2, 5
8. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02). pp. 311–318. Association for Computational Linguistics, USA (2002) 5
9. van Veldhuizen, V., Botha, V., Lu, C., Cesur, M.E., Lipman, K.G., de Jong, E.D., Horlings, H., Sanchez, C., Snoek, C., Mann, R., Marcus, E., Teuwen, J.: Foundation models in medical imaging: A review and outlook. arXiv preprint arXiv:2506.09095 (2025) 1
10. Wu, J., Wang, Y., Zhong, Z., Liao, W., Trayanova, N., Jiao, Z., Bai, H.X.: Vision–language foundation model for 3d medical imaging. npj Artificial Intelligence **1**, 17 (2025) 1, 2
11. Xin, Y., Ates, G.C., Gong, K., Shao, W.: Med3dvlm: An efficient vision-language model for 3d medical image analysis. arXiv preprint arXiv:2503.20047 (2025) 1, 2, 12
12. Yi, Z., Xiao, T., Albert, M.V.: A survey on multimodal large language models in radiology for report generation and visual question answering. Information **16**(2), 136 (2025) 2

Table 11. 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	Fig 3
Pre-processing	4
The dataset and evaluation metric section are presented	4
Environment setting table is provided	2
Training protocol table is provided	3
Ablation study	7
Limitation and future work are presented	Yes
Reference format is consistent.	Yes