Cervical-RG: Automated Cervical Cancer Report Generation from 3D Multi-sequence MRI via CoT-guided Hierarchical Experts

Hanwen Zhang[1,2*], Yu Long[1,2*], Yimeng Fan[1,2*], Yu Wang[3*], Zhaoyi Zhan[2], Sen Wang[2], Yuncheng Jiang[4,5], Rui Sun[4,5], Zheng Xing[6], Zhen Li[4,5], Xiaohui Duan[3]✉, Weibing Zhao[2]✉

[1] Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science Technology, Beijing Institute of Technology, China
[2] Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China
[3] Department of Radiology, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University
[4] FNii-Shenzhen, China [5] SSE, CUHK-Shenzhen, China
[6] College of Computer Science and Software Engineering, Shenzhen University
weibingzhao@smbu.edu.cn, duanxh5@mail.sysu.edu.cn

Abstract. Cervical cancer remains a leading cause of cancer-related mortality among females globally, with diagnosis primarily relying on multi-sequence magnetic resonance imaging (MRI). However, existing Multi-modal Large Language Models (MLLMs) struggle with processing 3D multi-sequence inputs due to high computational complexity and inefficient long-sequence modeling. To this end, we present **Cervical-RG**, which, to the best of our knowledge, this is the first framework that utilizes 3D multi-sequence MRI images for automated report generation. International Federation of Gynecology and Obstetrics (FIGO) staging, which plays a critical role in cervical cancer management, is also incorporated into the report. The workflow consists of (1) image diagnosis generation. (2) Chain of Thought (CoT)-guided FIGO staging with rationale, and (3) cross-stage consistency verification. Meanwhile, the entire pipeline simulates the collaborative diagnostic process of multi-disciplinary experts in clinical practice. Besides, we propose a novel model to handle multi-sequence inputs, comprising a volumetric multi-sequence encoder and a Mamba-Transformer hybrid decoder, which integrates global attention with selective state-space modeling to effectively handle long-range dependencies and spatial relationships. To validate our method, we curate **Cervical-MD**—a multi-modal dataset comprising 3,137 volumetrically aligned MRI-report pairs across five sequences (ADC, T1CA, T1CS, T2A, T2S), annotated by two radiologists. Experimental results demonstrate state-of-the-art performance in automated cervical cancer report generation. Our codes will be open-sourced soon.

Keywords: Cervical Cancer · Multi-Modal Large Language Models · Report Generation · Chain of Thought

* Equal contribution. ✉ Corresponding author.
Code available at: https://github.com/LongYu-LY/Cervical-RG

1 Introduction

Cervical cancer is among the most prevalent cancers in females worldwide, ranking as the fourth leading cause of cancer-related deaths globally. Early diagnosis is vital for improving cure rates. Automated image report generation technology can enhance diagnostic efficiency, minimize human errors, and deliver standardized reports for large-scale screenings. Doctors typically rely on multi-sequence MRI images for cervical cancer diagnosis, while due to scarcity of publicly available datasets, current research on multi-sequence images primarily focuses on classification [11, 12, 17] and segmentation [16, 18, 22] tasks, with limited exploration into high quality report generation.

In recent years, Multi-modal Large Language Models (MLLMs) have demonstrated outstanding performance in generating text that aligns closely with visual features, thus providing strong support for report generation tasks. Existing 2D MLLMs [14, 19, 25, 28] have proven to be powerful tools for report generation by fine-tuning on medical image-text pairs. However, they fail to fully leverage the rich spatial information in 3D MRI images, limiting their ability to accurately extract geometric priors such as tumor location, size, and other relevant characteristics. Recently, many impressive works, such as CT2Rep [7] and M3D-LaMed [3], have extended models to the 3D domain, achieving superior results. However, utilizing multi-sequence 3D MRI inputs for report generation has remained largely unexplored, mainly due to the absence of datasets and the substantial computational resources required for model training.

To this end, we have collected the **Cervical-MD** dataset, which includes 3,177 cases of cervical cancer patients from seven tertiary hospitals. The dataset consists of five 3D MRI sequences for each patient, along with their corresponding imaging diagnoses, and staging information labeled according to the FIGO 2018 criteria [6]. This is the first multi-modal dataset for cervical cancer. Additionally, we propose the **Cervical-RG** framework for report generation. Beyond generating image diagnosis, we also integrate FIGO staging prediction into the report, which provides more clinically relevant information and helps to identify potential high-risk patients. To enhance clinical interpretability, we divide the report generation process into three phases: image diagnosis generation, FIGO staging, and cross-stage consistency verification. In each phase, the domain-specific expert takes on a key role, mirroring the collaborative workflow in clinical practice. The radiologist specializes in generating imaging diagnoses, which entails producing descriptive findings from medical images and emulating the interpretation of tumor characteristics and associated abnormalities from multi-sequence imaging, guiding phase I. The oncologist, skilled in executing CoT and staging based on the available imaging features, leads phase II. A report reviewing specialist ensures the consistency and correctness of the staging reasoning in the phase III. To empower experts with the specialized skills required for analyzing multi-sequence images, we propose a novel model that includes a vision encoder specifically designed to extract 3D multi-sequence visual features and a hybrid Mamba-Transformer decoder to effectively balance computational efficiency and long-range dependency modeling.

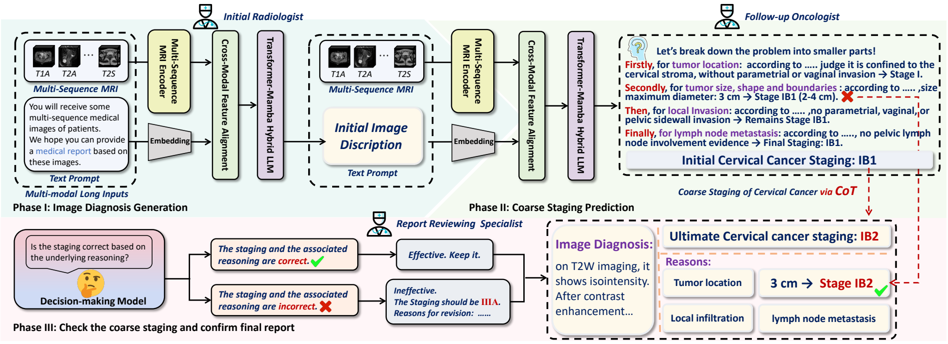

Initial Radiologist
T1A
T2A
T2S
Multi-Sequence MRI
You will receive some multi-sequence medical images of patients. We hope you can provide a medical report based on these images.
Text Prompt
Multi-modal Long Inputs
Multi-Sequence MRI Encoder
Embedding
Cross-Modal Feature Alignment
Transformer-Mamba Hybrid LLM
Phase I: Image Diagnosis Generation
T1A
T2A
T2S
Multi-Sequence MRI
Initial Image Discription
Text Prompt
Multi-Sequence MRI Encoder
Embedding
Cross-Modal Feature Alignment
Transformer-Mamba Hybrid LLM
Follow-up Oncologist
Let's break down the problem into smaller parts!
Firstly, for tumor location: according to ..... judge it is confined to the cervical stroma, without parametrial or vaginal invasion → Stage I.
Secondly, for tumor size, shape and boundaries : according to ..... ,size maximum diameter: 3 cm → Stage IB1 (2-4 cm).
Then, for local Invasion: according to ..... ,no parametrial, vaginal, or pelvic sidewall invasion → Remains Stage IB1.
Finally, for lymph node metastasis: according to ....., no pelvic lymph node involvement evidence → Final Staging: IB1.
Initial Cervical Cancer Staging: IB1
Coarse Staging of Cervical Cancer via CoT
Phase II: Coarse Staging Prediction
Report Reviewing Specialist
Is the staging correct based on the underlying reasoning?
Decision-making Model
The staging and the associated reasoning are correct.
The staging and the associated reasoning are incorrect.
Effective. Keep it.
Ineffective.
The Staging should be IIIA.
Reasons for revision: ......
Image Diagnosis:
on T2W imaging, it shows isointensity. After contrast enhancement...
Ultimate Cervical cancer staging: IB2
Reasons:
Tumor location
3 cm → Stage IB2
Local infiltration
lymph node metastasis
Phase III: Check the coarse staging and confirm final report

Fig. 1. Overview of the Cervical-RG framework for report generation, comprising image diagnosis generation, coarse staging prediction and cross-stage consistency verification to confirm the final report.

In summary, the main contributions can be summarized as follows:

– We propose a novel and pioneering framework, **Cervical-RG**, specifically designed for processing multi-sequence 3D MRI images, enabling across-sequence interactive learning to enhance report quality.
– We introduce the first integration of FIGO staging into the report generation process, achieved through Chain-of-Thought reasoning and cross-stage consistency verification, establishing a new paradigm for automated staging and reporting.
– We present **Cervical-MD**, a large-scale multi-modal dataset with 3,177 annotated cervical cancer MRI cases, enabling comprehensive training.

2 Method

We propose **Cervical-RG**, a hierarchical approach for automated report generation in cervical cancer. As illustrated in Fig. 1, in phase I and phase II, we leverage a radiologist and an oncologist to generate the imaging diagnosis and FIGO staging. In the final phase, we employ an external decision-making model to review the generated imaging report, ensuring the exclusion of errors and addressing any underlying logical inconsistencies.

2.1 Model Architecture

3D Vision Feature Extraction. Due to the scarcity of publicly available MRI datasets, existing pre-trained 3D vision encoders struggle to effectively extract information from cervical MRI images. To this end, we employed GPT-4o [1] to extract an MRI-image-to-text alignment dataset from Cervical-MD. Specifically, it performs a professional parsing of the full radiology reports, breaking them down into diagnostic descriptions that correspond precisely to each MRI

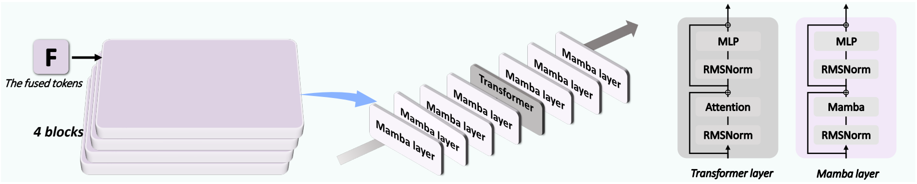


Fig. 2. Schematic of the Mamba-Transformer hybrid decoder architecture, consisting of four blocks, each containing eight layers.

sequence (e.g., ADC, T2A). After extraction, we review the image-text pairs for accuracy and consistency. The validated dataset, consisting of knowledge-enhanced image-text pairs $\{(I_i, T_i)\}_{i=1}^{5}$, is then used to pre-train our multi-sequence MRI encoder (based on ViT) with a CLIP-like approach. During this process, GPT-4o only plays the role of constructing the dataset.

Initially, we concatenate the five images along the first dimension, obtaining a tensor $\mathbf{I} \in \mathbb{R}^{5\times1\times32\times256\times256}$, with each image having a single channel and a spatial resolution of $32 \times 256 \times 256$. This tensor is then processed by our multi-sequence MRI encoder, which performs a patch embedding operation. Specifically, the images are divided into patches of size $4 \times 16 \times 16$, resulting in a flattened tensor $\mathbf{X} \in \mathbb{R}^{5\times2048\times768}$, where 2048 (i.e., $\frac{32}{4} \times \frac{256}{16} \times \frac{256}{16}$) denotes the total number of patches, and 768 represents the feature dimension after embedding.

Multi-Modal Feature Fusion. Inspired by [3], We reconstruct the output tokens $\mathbf{X}$ into a 3D tensor $\mathbf{X_{reconstructed}} \in \mathbb{R}^{5\times\frac{32}{4}\times\frac{256}{16}\times\frac{256}{16}\times768}$. A 3D-spatial pooling operation (size = 2) is then applied to reduce the number of tokens, producing $\mathbf{X_{pooled}} \in \mathbb{R}^{5\times\frac{8}{2}\times\frac{16}{2}\times\frac{16}{2}\times768}$, This step reduces computational cost while maintaining essential spatial features. Following pooling, a projector comprising Multi-Layer Perceptrons (MLPs) used to adjust the embedding dimensions of the tokens, ensures alignment with decoder and then obtain $\mathbf{Z} = \mathrm{MLP}(\mathbf{X_{pooled}}) \in \mathbb{R}^{5\times(\frac{8}{2}\times\frac{16}{2}\times\frac{16}{2})\times768}$. After that, we split this tensor along the first dimension to obtain the visual tokens corresponding to each MRI sequence. For text processing, we use SentencePiece [13] to tokenize the text, and visual tokens are then inserted into predefined positions of image tags (`"<IMAGE>"`) within the text tokens, enabling fusion of visual and textual features as $\mathbf{F}$.

Mamba-Transformer hybrid Decoder. As is shown in Fig. 2, we employ a hybrid decoder architecture consisting of four blocks of hybrid layers to decode the fused tokens $\mathbf{F}$. Each block is composed of Mamba and Transformer layers, as observed in [23] the most significant improvement was observed while the ratio transitioned from 1:0 to 7:1, so we chose the 7:1 ratio. To stabilize training at large model scales, RMSNorm [26] is applied within the layers. After decoding, the fused tokens $\mathbf{F}$ yield the generated text $\mathbf{G}$, the model's final output.

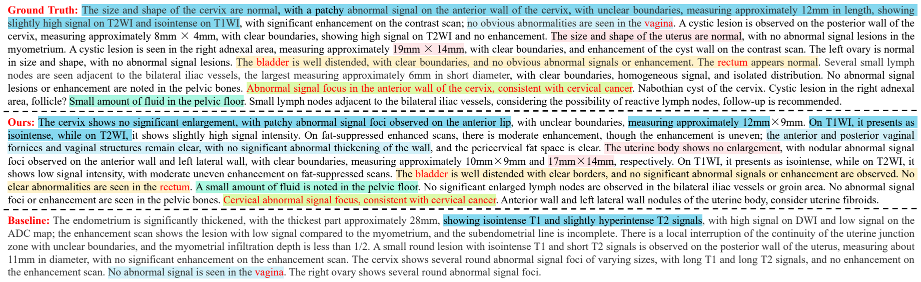
2.2 Hierarchical Experts Mechanism

Inspired by the multi-disciplinary collaboration process in clinical consultations, we propose a CoT-guided hierarchical experts mechanism for report generation. In this approach, the initial **radiologist** acts as the expert in imaging diagnosis, the follow-up **oncologist** generates the initial FIGO staging, and the **report reviewing specialist** ensures the quality of the final report. The entire hierarchical process not only generates accurate reports but also ensures the interpretability of the reasoning process. Our final report consists of the image diagnosis, FIGO staging, and the corresponding correct reasoning process.

Initial Radiologist. We fine-tune our model on multi-sequence images and their corresponding image diagnoses from Cervical-MD, training it to simulate the role of an radiologist in generating image-based diagnosis. The model learns to describe lesions by integrating visual cues from medical images and contextual information from corresponding text, producing accurate, clinically relevant diagnosis aligned with the image data. To train the model, we employ Cross-Entropy Loss function, defined as $\mathcal{L} = -\sum_{t=1}^{T}\sum_{i=1}^{V} y_{t,i}\log(p_{t,i})$, where T is the length of the text sequence, and V is the vocabulary size. The term $y_{t,i}$ represents the actual label for generating the i-th word in the vocabulary at the t-th time step. Conversely, $p_{t,i}$ refers to the probability that the model predicts for generating the i-th word in the vocabulary at the t-th time step.

Follow-up Oncologist. During the clinical staging phase, oncologists must synthesize radiological findings with FIGO staging criteria through systematic clinical reasoning. To emulate this complex decision-making process, we propose a structured reasoning framework that decomposes cervical cancer staging into clinically interpretable sub-tasks corresponding to key FIGO parameters. Specifically, we introduce six reasoning tags: `<SIZE>`, `<LOCATION>`, `<INFILTRATE>`, `<INVOLVEMENT>`, `<OTHER>`, and `<STAGING>`. During data construction, each indicator is annotated using these structured tags to explicitly mark the beginning and end of its reasoning segment (e.g., <SIZE>The tumor size is 34mm*12mm*10mm</SIZE>). These labels correspond to the reasoning process, including tumor size, location, local infiltration, pelvic wall involvement, as well as descriptions of other lesions and the final staging results, guiding the model in making staging decisions based on the image diagnosis during training. We also utilize the Cross-Entropy Loss function for model training, consistent with the approach described in the previous section. This hierarchical annotation strategy provides explicit supervision signals during fine-tuning, enabling the model to learn decision pathways that mirror oncologists' diagnostic workflows. By incorporating Chain-of-Thought (CoT) reasoning, the model can simulate the step-by-step logical process followed by an oncologist to determine the staging decisions. This not only improves the model's ability to reason through complex medical tasks but also ensures the auditability and interpretability of its decision-making process.

Report Reviewing Specialist. Due to the potential logical inconsistencies between image diagnosis and FIGO staging in the text generated by the MLLM, we



Ground Truth: The size and shape of the cervix are normal, with a patchy abnormal signal on the anterior wall of the cervix, with unclear boundaries, measuring approximately 12mm in length, showing slightly high signal on T2WI and isointense on T1WI, with significant enhancement on the contrast scan; no obvious abnormalities are seen in the vagina. A cystic lesion is observed on the posterior wall of the cervix, measuring approximately 8mm × 4mm, with clear boundaries, showing high signal on T2WI and no enhancement. The size and shape of the uterus are normal, with no abnormal signal lesions in the myometrium. A cystic lesion is seen in the right adnexal area, measuring approximately 19mm × 14mm, with clear boundaries, and enhancement of the cyst wall on the contrast scan. The left ovary is normal in size and shape, with no abnormal signal lesions. The bladder is well distended, with clear boundaries, and no obvious abnormal signals or enhancement. The rectum appears normal. Several small lymph nodes are seen adjacent to the bilateral iliac vessels, the largest measuring approximately 6mm in short diameter, with clear boundaries, homogeneous signal, and isolated distribution. No abnormal signal lesions or enhancement are noted in the pelvic bones. Abnormal signal focus in the anterior wall of the cervix, consistent with cervical cancer. Nabothian cyst of the cervix. Cystic lesion in the right adnexal area, follicle? Small amount of fluid in the pelvic floor. Small lymph nodes adjacent to the bilateral iliac vessels, considering the possibility of reactive lymph nodes, follow-up is recommended.

Ours: The cervix shows no significant enlargement, with patchy abnormal signal foci observed on the anterior lip, with unclear boundaries, measuring approximately 12mm×9mm. On T1WI, it presents as isointense, while on T2WI, it shows slightly high signal intensity. On fat-suppressed enhanced scans, there is moderate enhancement, though the enhancement is uneven; the anterior and posterior vaginal fornices and vaginal structures remain clear, with no significant abnormal thickening of the wall, and the pericervical fat space is clear. The uterine body shows no enlargement, with nodular abnormal signal foci observed on the anterior wall and left lateral wall, with clear boundaries, measuring approximately 10mm×9mm and 17mm×14mm, respectively. On T1WI, it presents as isointense, while on T2WI, it shows low signal intensity, with moderate uneven enhancement on fat-suppressed scans. The bladder is well distended with clear borders, and no significant abnormal signals or enhancement are observed. No clear abnormalities are seen in the rectum. A small amount of fluid is noted in the pelvic floor. No significant enlarged lymph nodes are observed in the bilateral iliac vessels or groin area. No abnormal signal foci or enhancement are seen in the pelvic bones. Cervical abnormal signal focus, consistent with cervical cancer. Anterior wall and left lateral wall nodules of the uterine body, consider uterine fibroids.

Baseline: The endometrium is significantly thickened, with the thickest part approximately 28mm, showing isointense T1 and slightly hyperintense T2 signals, with high signal on DWI and low signal on the ADC map; the enhancement scan shows the lesion with low signal compared to the myometrium, and the subendometrial line is incomplete. There is a local interruption of the continuity of the uterine junction zone with unclear boundaries, and the myometrial infiltration depth is less than 1/2. A small round lesion with isointense T1 and short T2 signals is observed on the posterior wall of the uterus, measuring about 11mm in diameter, with no significant enhancement on the enhancement scan. The cervix shows several round abnormal signal foci of varying sizes, with long T1 and long T2 signals, and no enhancement on the enhancement scan. No abnormal signal is seen in the vagina. The right ovary shows several round abnormal signal foci.

Fig. 3. Comparison of ground-truth with image diagnoses generated by our model and Baseline model (M3D-LaMed). For better illustration, The same color corresponds to the same expression medical term.

introduce an external decision-making model as report reviewing specialist for cross-stage consistency verification. Specifically, we input image diagnosis, staging criteria, and initial staging result into the model, which performs logical checks to identify any potential inconsistencies or errors. If the model detects any issues with the staging judgment, it will automatically correct them, thereby preventing conflicts between the initial FIGO result and staging criteria and ensuring that the generated text aligns more closely with clinical reality.

3 Experiments and Results

3.1 Datasets

We present the first large-scale, multi-sequence MRI dataset for cervical cancer, named **Cervical-MD**. This dataset encompasses medical data from 3,177 patients, collected across seven tertiary hospitals. The patients range in age from 21 to 90 years, with the 41–50 and 51–60 age groups being the most prevalent, accounting for 30.16% and 38.32% of the cases, respectively. Each case contains five MRI sequences (ADC, T1CA, T1CS, T2A, T2S), along with detailed imaging conclusions and the FIGO stage recorded by clinicians. To ensure statistical validity under limited data availability, 15 cases were randomly selected from each hospital to form the test set of 105 test cases. The remaining cases were utilized for training.

3.2 Implementation Details

In the experiments for phase I and II, we used the same experimental setup. Specifically, we employed LongLLaVA-med [14] as the pre-trained weights for our Mamba-Transformer hybrid decoder. We utilized a two-stage training strategy for our model. First, we froze the parameters of both the vision encoder and the decoder, training only the projector. Then, we froze the vision encoder and fine-tuned the projector and LLM with full parameter updates. For both stages, we set the learning rate to $1e^{-5}$, batch size to 1, and the number of epochs to 6. The training was conducted on 8 A6000 GPUs 48 GB with the Zero2 technique. In Phase III, we employed DeepSeek-R1-Distill-Llama-8B [5], which supports local deployment, as our external decision-making

Table 1. Image Diagnosis generation performance comparison of various models on Cervical-MD dataset. † denotes models that have been retrained using the Cervical-MD dataset. The best results are highlighted in **bold**.

Model	BLEU-2 ↑	ROUGE-1 ↑	METEOR ↑	Bert-Score ↑	RadGraph ↑	RadCliQ ↓
LLaVA-Med [14]	0.021	0.158	0.079	0.812	0.012	2.335
GPT-4o [1]	0.029	0.151	0.102	0.804	0.021	2.391
R2GenGPT† [25]	0.167	0.358	0.247	0.841	0.149	1.768
miniGPT-Med† [2]	0.231	0.492	0.330	0.880	0.261	1.221
LongLLaVA-Med† [24]	0.253	0.504	0.358	0.882	0.283	1.157
M3D-LaMed† [3]	0.191	0.432	0.299	0.867	0.218	1.427
Cervical-RG (Ours)	**0.286**	**0.536**	**0.394**	**0.888**	**0.297**	**1.068**

model, which has been tested to possess the ability to discriminate according to rules and description.

3.3 Evaluation Metrics

To comprehensively evaluate the generated reports, we assessed them from two perspectives. (1) Using traditional natural language processing metrics BLEU-2 [20], ROUGE-L [15], METEOR [4], and BertScore [27] alongside more recent medical report evaluation metrics, including RadGraph F1 [10], and RadCliQ [10], to assess imaging diagnoses. (2) Using accuracy (Acc) to evaluate FIGO staging prediction, with C-Acc for coarse-grained stages (e.g., I, II, III, IV) and F-Acc for fine-grained stages (e.g., IA, IIA1). Additionally, we conducted a clinician evaluation with two experienced doctors scoring each report from 1 to 10 based on staging accuracy (0-4), logical coherence (0-4), and clinical evidence adequacy (0-2). To mitigate subjective variability, a standardized scoring rubric was employed by both clinicians to ensure inter-rater consistency.

3.4 Comparison with the Baseline Method

Image Diagnosis Generation. We compared the performance of our approach with recent 2D and 3D medical MLLMs, as shown in Table 1. For LLaVA-Med [14], GPT-4o [1], we used their official interfaces and checkpoint for zero-shot evaluation. Additionally, for miniGPT-Med [2], R2GenGPT [25], LongLLaVA-Med [24], and M3D-LaMed [3], we utilized their publicly released training code to fine-tune them on the Cervical-MD training set, followed by evaluation on the test set. Since existing models are unable to handle multi-sequence 3D inputs, we conducted the tests using the model with the highest input configuration requirements. The results highlight that our model outperforms competing approaches, underscoring the superiority of leveraging 3D multi-sequence image inputs for generating more accurate imaging diagnosis, Fig. 3 shows an example of generated image diagnosis.

FIGO Stage Results. We formulate FIGO staging as a classification task, employing both UNETR [9] and Swin-UNETR [8] architectures for stage prediction. For C-Acc, we treat it as a 4-class task, and for F-Acc, as a 19-class task based on our dataset. As illustrated in the left panel of Table 2, conventional classification frameworks may be suboptimal for staging tasks due to their heightened sensitivity to nuanced imaging features. To address this complexity, our model implements a chain-of-thought reasoning paradigm that systematically decomposes the diagnostic process, ultimately

Table 2. Comparison of our method with UNETR [9] and Swin-UNETR [8] for classification tasks (left) and ablation results of different expert stages (right). Expert-1 to Expert-3 correspond to the Radiologist, Oncologist, and Report Reviewing Specialist.

Model	C-Acc ↑	F-Acc ↑	Expert-1	Expert-2	Expert-3	C-Acc ↑	F-Acc ↑	Score ↑
UNETR [9]	0.283	0.124	✓		✓	0.352	0.124	/
Swin-UNETR [8]	0.333	0.133	✓	✓		0.543	0.181	4.474
Cervical-RG (Ours)	**0.642**	**0.264**	✓	✓	✓	**0.642**	**0.264**	**6.320**

Table 3. Ablation experiment results in this table focusing on comparisons of different inputs (2D vs. 3D, number of sequences) in the upper section, and the influence of retraining the visual encoder and projector in the lower section.

Comparing Different Input Sequence							
	Sequence	BLEU-2	ROUGE-1	METEOR	Bert-Score	RadGraph	RadCliQ
2D	T2A	0.244	0.497	0.350	0.881	0.272	1.197
3D	T2A	0.260	0.507	0.359	0.882	0.222	1.196
2D	ADC,T1CS,T2A	0.252	0.504	0.353	0.881	0.277	1.179
3D	ADC,T1CS,T2A	0.259	0.504	0.364	0.882	0.277	1.157
2D	Full Sequences	0.253	0.504	0.358	0.882	0.283	1.157
3D	Full Sequences	**0.286**	**0.536**	**0.394**	**0.888**	**0.297**	**1.068**
Effect of Multi-Stage Training Strategy							
Encoder	Projector	BLEU-2	ROUGE-1	METEOR	Bert-Score	RadGraph	RadCliQ
		0.211	0.454	0.330	0.868	0.207	1.446
✓		0.270	0.520	0.376	0.885	0.295	1.077
✓	✓	**0.286**	**0.536**	**0.394**	**0.888**	**0.297**	**1.068**

demonstrating state-of-the-art performance in staging accuracy. Although the F-Acc has not yet achieved a high value due to the complexity of staging tasks and the inability of MRI images to provide all the information required for FIGO staging, clinicians consider our results to have met the minimum diagnostic acceptability criteria.

3.5 Ablation Studies

Comparison of 2D vs. 3D Inputs and Multi-sequence Impact. As shown in Table 3, our 3D visual encoder outperforms the 2D ViT-L/14 [21] across all metrics, validating the superior diagnostic capability of volumetric MRI. Performance improves with the inclusion of more input sequences, demonstrating the complementarity and synergistic effect of multi-sequence data in enhancing diagnostic accuracy.

Multi-Stage Training Strategy. Considering that pre-trained components may not have been trained on 3D MRI data, potentially hindering feature alignment and inference, we adopted a multi-stage training strategy. As demonstrated in the lower section of Table 3, the performance improves progressively with each stage of feature alignment. This indicates that current MLLMs still face challenges in effectively extracting and aligning features from 3D MRI data.

Effectiveness of Hierarchical Experts for FIGO staging. To evaluate the impact of the hierarchical experts framework on FIGO staging, we conducted three experiments based on the image diagnosis generated by the Radiologist: (1) directly put the image diagnosis into the external decision-making model for staging; (2) with the oncologist determine the staging based on the diagnosis; and (3) the report review specialist refine the staging made by oncologist according image diagnosis. The results, shown in the right side of Table 2, highlight the necessity of each expert and validate the effectiveness of the CoT-guided approach and cross-stage consistency verification.

4 Conclusion

We introduce Cervical-RG, a novel framework for cervical cancer report generation that efficiently handles long 3D multi-sequence MRI inputs under constrained computational resources, enabled by a hybrid Mamba-Transformer architecture. By deeply integrating multi-sequence data and incorporating FIGO staging with hierarchical experts mechanism, our approach generates accurate and clinically reliable reports. Extensive experiments and clinical validations by doctors validate its superior performance, establishing a robust solution for automated cervical cancer diagnosis.

Acknowledgments. This work was partially supported by the Shenzhen Science and Technology Program (Grant No. JCYJ20241202130548062), and the Guangdong Provincial Key Area Project of General Universities (Grant No. 2024ZDZX1017).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
2. Asma Alkhaldi, Raneem Alnajim, Layan Alabdullatef, Rawan Alyahya, Jun Chen, Deyao Zhu, Ahmed Alsinan, and Mohamed Elhoseiny. Minigpt-med: Large language model as a general interface for radiology diagnosis. *arXiv preprint arXiv:2407.04106*, 2024.
3. Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.
4. Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
5. Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.
6. Perry W Grigsby, Leslie S Massad, David G Mutch, Matthew A Powell, Premal H Thaker, Carolyn McCourt, Andrea Hagemann, Katherine Fuh, Lindsay Kuroki, Julie K Schwarz, et al. Figo 2018 staging criteria for cervical cancer: Impact on stage migration and survival. *Gynecologic oncology*, 157(3):639–643, 2020.
7. Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–486. Springer, 2024.

8. Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.
9. Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.
10. Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.
11. Xiran Jiang, Jiaxin Li, Yangyang Kan, Tao Yu, Shijie Chang, Xianzheng Sha, Hairong Zheng, Yahong Luo, and Shanshan Wang. Mri based radiomics approach with deep learning for prediction of vessel invasion in early-stage cervical cancer. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(3):995–1002, 2020.
12. Smith K Khare, Berit Bargum Booth, Victoria Blanes-Vidal, Lone Kjeld Petersen, and Esmaeil S Nadimi. An explainable attention model for cervical precancer risk classification using colposcopic images. *arXiv preprint arXiv:2411.09469*, 2024.
13. Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018.
14. Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.
15. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
16. Yu-Chun Lin, Yenpo Lin, Yen-Ling Huang, Chih-Yi Ho, Hsin-Ju Chiang, Hsin-Ying Lu, Chun-Chieh Wang, Jiun-Jie Wang, Shu-Hang Ng, Chyong-Huey Lai, et al. Generalizable transfer learning of automated tumor segmentation from cervical cancers toward a universal model for uterine malignancies in diffusion-weighted mri. *Insights into Imaging*, 14(1):14, 2023.
17. Shuyu Liu, Yu Zhou, Caizhi Wang, Junjie Shen, and Yi Zheng. Prediction of lymph node status in patients with early-stage cervical cancer based on radiomic features of magnetic resonance imaging (mri) images. *BMC Medical Imaging*, 23(1):101, 2023.
18. Pengyue Lu, Faming Fang, He Zhang, Lei Ling, and Keqin Hua. Augms-net: Augmented multiscale network for small cervical tumor segmentation from mri volumes. *Computers in Biology and Medicine*, 141:104774, 2022.
19. Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.
20. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

21. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.
22. Awj Twam, Megan Jacobsen, Rachel Glenn, Ann Klopp, Aradhana M Venkatesan, and David Fuentes. Two stage segmentation of cervical tumors using pocketnet. *arXiv preprint arXiv:2409.11456*, 2024.
23. Junxiong Wang, Daniele Paliotta, Avner May, Alexander Rush, and Tri Dao. The mamba in the llama: Distilling and accelerating hybrid models. *Advances in Neural Information Processing Systems*, 37:62432–62457, 2025.
24. Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2024.
25. Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023.
26. Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.
27. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
28. Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.