PRG-RGen: Pathologist Reasoning-Guided Report Generation on WSIs via Multimodal LLMs

Satoshi Kondo[1] Satoshi Kasai[2]

[1] Muroran Institute of Technology, Hokkaido, Japan
[2] Fujita Health University, Aichi, Japan
kondo@muroran-it.ac.jp

Abstract. A key challenge in the automatic generation of pathology reports from gigapixel WSI images is the black-box nature of the reasoning process. In this study, in accordance with the MICCAI 2026 REG² Challenge, we propose a two-module framework that mimics the thought process of a pathologist. The reasoning module uses a minimum path coverage algorithm to extract and learn optimal diagnostic pathways from a DAG. The grounding module performs region-of-interest (ROI) identification that is robust to image perturbations through binary classification using consistency loss. The overall score for Test Phase 1 in the official challenge system is 0.5444. The code is available at https://github.com/sk-jp/REG2026.

Keywords: Pathology, Chain-of-Thought Reasoning, Visual Grounding

1 Introduction

In recent years, advancements in deep learning and multimodal large-scale language models (MLLM) have led to dramatic progress in technologies that automatically generate diagnostic reports from gigapixel-scale whole-slide images (WSI). However, conventional evaluation metrics rely on superficial lexical matching, such as ROUGE-L and BLEU-4, and thus present the following two challenges:

1. Lack of clinical validity: Simple word matching cannot evaluate clinical errors such as misinterpretation of negative expressions or omission of important findings.
2. Opacity of the reasoning process: Black-box reasoning struggles to gain the trust of physicians and has low acceptance in clinical settings.

To address these challenges, this study proposes a new report generation framework that mimics the pathologist's thought process (Chain-of-Thought: CoT), in accordance with the REG² Challenge at MICCAI 2026 [1, 2]. Specifically, it simultaneously solves the following two tasks:

1. Workflow Reasoning: Outputs the pathological rationale leading to the diagnosis step-by-step in structured language.
2. Visual Grounding: Outputs the location information of regions of interest (ROIs) on the WSI that serves as evidence for key findings in the report.

This enables healthcare professionals to audit and verify the AI's reasoning, thereby improving diagnostic safety.

2 Related Works

2.1 Automatic Pathology Report Generation for WSI

In recent years, the emergence of pathology image-based models, such as UNI [3], has made it possible to represent the vast spatial context of WSI. Furthermore, multimodal learning approaches that combine vision and language, such as MI-Zero [4], have enabled zero-shot alignment between patch images and diagnostic text, dramatically improving report quality. However, these methods rely on end-to-end learning that generates reports directly from WSI, and a key challenge is that the reasoning process leading to the final decision remains a complete black box.

2.2 Multimodal Chain of Thought (CoT) in Medical Imaging

Efforts to integrate CoT into medical image diagnosis are gaining momentum, as seen in Med-CoT [5], which sequentially generates clinical decision-making paths, and MedEyes [6], which uses reinforcement learning to model a physician's dynamic eye movements and thought processes. However, these approaches typically target local images of standard size (e.g., 512×512 pixels) or radiological images, and their extension to the "slide-exploration-based reasoning workflow", which involves traversing vast fields of view in pathological WSI, has yet to be established.

2.3 Visual Grounding in Medical Images

Visual grounding, the process of spatially identifying image regions corresponding to descriptions in a report, is crucial for ensuring the reliability of the text. RadGrounder [7] and other studies have demonstrated the effectiveness of multi-task learning that simultaneously performs report generation and region prediction. This approach is also essential in light of findings from studies such as VGMED [8], which pointed out a phenomenon where medical MLLMs output correct answers regarding text strings (semantic grounding) while focusing on incorrect image regions (failure of visual grounding). However, a technology capable of pinpointing minute ROIs within vast WSI images and seamlessly linking them to pathological reasons has not yet been established.

3 Proposed Method

To address the two evaluation criteria, stepwise differential diagnosis reasoning for whole-slide images (WSI) and grounding through tissue/background classification at the region of interest (ROI) level, our proposed method consists of two independently trained modules, the Reasoning module (CONCH encoder [9] + Qwen3.5-2B [10]) and the Grounding module (PLIP encoder [11] + Qwen2.5-7B [12]). The common design

principle for both modules is to map visual features obtained from a frozen pathology image base model to the token space of a large language model (LLM) via a lightweight projection mechanism. The LLM then uses QLoRA for low-rank adaptation, thereby specializing it for pathology diagnosis tasks without the need to retrain a large-scale image encoder. Furthermore, the evaluation metrics themselves are incorporated into the training process, threshold determination.

3.1 Reasoning Module

Training Data Construction.

The training data for the reasoning module was constructed using only deterministic algorithms based on the correct answer annotations for the chain-of-thought (CoT) queries assigned to the dataset. The CoT annotations assigned to each case are provided as a set of transitions consisting of a question, an answer, and a follow-up question, and could include multiple branches within the same case (a structure where multiple follow-up questions derive from the same question-answer pair). Therefore, we first consolidate identical question-answer pairs into a single node to construct a directed acyclic graph (DAG) consisting solely of forward edges based on the order of occurrence. Next, we enumerate all paths from the start node to the end node on this DAG and, by performing a greedy selection in descending order of path length, determine the set of paths with the minimum number required to cover all nodes in the graph (minimum path coverage). Through this process, even when a single case involves multiple differential diagnoses or branching reasoning processes, the reasoning structure inherent in the case can be converted into training samples without over- or under-representation, and without including redundant sub-paths in the training data. Each obtained path is formatted into a single training dialogue as a series of questions and answers ranging from “organ identification” to “the description in the final pathology report,” and is assigned image tokens and fixed instructions (directing step-by-step reasoning and the generation of the final pathology report).

For the image-side input representation, patch features extracted from each WSI using the pathology image foundation model (CONCH) are pre-computed and cached. Patch selection do not rely on simple uniform random sampling; instead, we employ tissue-priority sampling, which scores candidate patch sets based on tissue area coverage and saturation, and selects the top-ranked patches. This selection policy aligns with the patch sampling method used during final reasoning, with the intent of minimizing the divergence between the distributions used in training and reasoning. For each patch, we also retain its normalized coordinates relative to the entire slide, which are used for the spatial position encoding described later.

Model Architecture.

The reasoning module first encodes each patch using the frozen pathological image base model CONCH (Vision Transformer, 1024-dimensional), then applies a sinusoidal 2D positional encoding based on normalized slide coordinates to each patch feature, thereby explicitly incorporating information about the spatial arrangement on the slide.

The set of patch features with the added positional encoding is contextualized by a Transformer encoder based on a self-attention mechanism and then aggregated by the subsequent Perceiver Re-sampler into a fixed-length sequence of visual tokens (16 in our implementation) that is independent of the number of patches. This sequence of visual tokens is mapped to the embedding space of the language model via a multi-layer perceptron projection, concatenated to the beginning of the text token sequence, and then fed into a single autoregressive language model.

We use the Qwen 3.5 series for the language model; the final configuration adopted is a model with 2B parameters, and under 4-bit quantization, low-rank adaptation via LoRA is applied to all attention projection layers and feedforward layers.

Loss Function and Model Selection.

The loss function used during training is an autoregressive language modeling loss (cross-entropy loss) for the token sequence, excluding the losses for visual tokens and tokens corresponding to instructions. Here, to prioritize the accurate generation of the final pathology diagnosis report over the step-by-step reasoning process, we use weighted cross-entropy loss, which assigns a larger loss weight to the token intervals corresponding to the final diagnosis report than to other tokens.

Furthermore, at the end of each training epoch, we perform actual autoregressive generation on a subset of the validation cases. We then reconstruct the resulting CoT sequences into structured question-answer pairs and directly calculate the official evaluation metrics for this task (a composite metric including scores based on the validity of the reasoning path and the lexical and semantic similarity of the final report). For model selection, we maintain checkpoints based on both validation loss and this official evaluation metric, thereby mitigating the discrepancy between a decrease in training loss and an improvement in actual downstream evaluation performance.

3.2 Grounding Module

Construction of Training Data.

The grounding module is formulated as a task to determine whether a region of interest (ROI) on a WSI contains tissue. The training data is constructed by extracting fixed-size ROI images from the WSI and accompanying pathological diagnosis information for each case, converting the magnification from high to low (in practice, from a slide equivalent to 20x magnification to a field of view equivalent to 5x magnification). For each WSI thumbnail image, a tissue region mask is derived by combining color-space-based thresholding with morphological processing. Based on this mask, three types of ROIs are extracted from each case: (i) ROIs containing tissue, (ii) background ROIs not containing tissue, and (iii) background ROIs (hard negatives) taken from the vicinity of the boundary between the tissue and background regions, where distinction is difficult. A common question (asking whether tissue is visible within the ROI in question) is assigned to all ROIs, and the response is set to one of two standard phrases depending on whether the ROI contained tissue or background. The reason for deliberately standardizing the responses is to ensure semantically stable reference

sentences for the same ROI when evaluating and learning the consistency of responses to perturbations, as described later.

Model Architecture.

The grounding module encodes ROI images using the frozen pathological image base model PLIP. It then maps the resulting image features into multiple (8) visual tokens using a multilayer perceptron and inputs these tokens into a language model (Qwen2.5-7B-Instruct) that has undergone 4-bit quantization and low-rank adaptation via LoRA. In addition, a lightweight linear classification head that acts directly on PLIP's raw image features is implemented in parallel to perform a binary classification of whether the ROI contains tissue.

Loss Function and Model Selection.

During training, pairs of images are generated for each ROI by applying perturbations such as hue, contrast, blurring, JPEG recompression, and partial occlusion; both the original and perturbed images are then fed into the classification head. The overall training objective function is constructed as a weighted sum of: (i) autoregressive cross-entropy loss for response generation by the language model; (ii) cross-entropy loss for binary classification of the original and perturbed images, respectively; and (iii) consistency loss (mean squared error) that penalizes the discrepancy between the classification probabilities of the original and perturbed images. By introducing this consistency loss, we explicitly incorporate into the training the requirement that the classification results be robust against image quality degradation and scanner-induced artifacts.

For evaluation, we directly calculate the official evaluation metric for this task within the training loop as a weighted sum of three components: the correct rejection rate of background regions, the agreement rate between pre- and post-perturbation classifications, and the simultaneous correct classification rate of tissue-background pairs within the same case. We then determine the threshold that maximizes this metric by exhaustively searching for classification thresholds on the validation set at each epoch. The final reasoning is performed using a lightweight architecture that do not involve a language model, relying solely on the image encoder, a lightweight classification head, and the threshold determined during training.

4 Experiments and Results

4.1 Dataset

The cases used in this experiment consisted of full-slide images from a total of 11,220 cases spanning seven organs (primarily the prostate, breast, stomach, colon, bladder, lung, and cervix). Data splitting was performed at the case (WSI) level, and under a fixed random seed, the dataset was divided into 8,976 cases for training and 2,244 cases for evaluation (approximately 80:20). This case-level splitting was used

consistently for both the reasoning module and the grounding module and was configured so that no single case appeared in both the training and evaluation sets.

Data for Reasoning Module.

By expanding the step-by-step reasoning paths extracted from each case (1 to 10 paths per case using the minimum path coverage algorithm described in Section 3.1) into training dialogues, we obtained 35,258 samples for training and 8,769 samples for evaluation. Table 1 shows the number of cases by organ (counted as 1 case per instance before path expansion).

Table 1. Organ distribution of data used for reasoning (by case)

Organ	Train (n=8,976)	Validation (n=2,244)
Prostate	1,949	469
Breast	1,770	442
Stomach	1,406	351
Colon	1,380	371
Bladder	866	213
Lung	755	190
Cervix	647	165
Rectum	201	43

Data for Grounding Module.

For data splitting, the data was stratified by combinations of organ and tumor type and then divided into training and validation sets. Based on the same case segmentation, the number of regions of interest (ROIs) extracted from each WSI was 85,177 for training (corresponding to 8,968 WSIs) and 21,275 for evaluation (corresponding to 2,243 WSIs). Table 2 shows the breakdown of these ROIs into those containing tissue and background ROIs (including hard negatives near the boundaries). Furthermore, for each organ, we adopted a method of extracting a fixed number of ROIs per case to ensure that the number of cases was proportional to the total number of cases.

Table 2. Number of ROI data for grounding

ROI	Train	Validation
Background	53,504	13,379
(Background near the boundary)	(17,831)	(4,459)
Tissue	31,673	7,896

4.2 Experimental Conditions

Reasoning Module.

The training conditions for the reasoning module are described below. We used a frozen version of CONCH v1.5 as the base image encoder, and adapted Qwen3.5-2B for the language model using QLoRA with 4-bit quantization. LoRA was applied to all projection layers corresponding to the attention mechanism and gated linear attention mechanism, as well as to all feed forward layers. The rank, alpha and dropout rate in LoRA was 16, 32 and 0.05, respectively. Number of sample patches per case was 32, the maximum sequence length was 768 tokens, the loss weight for the final reporting interval was 2.0, the batch size was 8, the number of epochs was 20, and the learning rate scheduler was cosine annealing with warm-up (warm-up ratio 0.1).

Grounding Module.

The training conditions for the grounding module are described below. We used a frozen PLIP model as the image encoder and Qwen2.5-7B-Instruct as the language model, adapted using 4-bit QLoRA. The rank, alpha and dropout rate in LoRA was 16, 32 and 0.05, respectively. The number of image tokens was 8. The loss weights for response generation, classification and consistency was 1.0, 0.5 and 0.5, respectively. The probability of masking and masking area ratio during perturbation are 0.3 and 15-25 % of image area, respectively. The maximum sequence length was 768 tokens, the batch size was 128, the number of epochs was 10, and the learning rate scheduler was cosine annealing with warm-up (warm-up ratio 0.1).

4.3 Results

Reasoning Module.

We performed a grid search for the learning rate in the range of 1×10^{-5} to 1×10^{-4} and selected the learning rate that yielded the best evaluation score. Evaluation was performed at the end of each epoch by generating actual reasoning paths for a subset of the validation cases and calculating metrics based on the official evaluation scheme ('binary_path_validity', 'edge_f1', 'final_report_score_approx') and composite metric ('workflow_ranking_score'). As a result, a learning rate of 6×10^{-5} was adopted as the final model.

Grounding Module.

Among the settings tested during the learning rate search, the setting using 6×10^{-3} yielded the highest value (0.982) for the evaluation ROI accuracy within the search range. The grounding module achieved a metric score of around 0.98 on the evaluation data and demonstrated high robustness, particularly in rejecting background regions and maintaining classification consistency under perturbations. However, the simultaneous positive classification rate for tissue-background pairs within the same case was somewhat low, suggesting that this is a major cause of misclassification in this module.

Results for Test Phase 1.

The results for Test Phase 1 in the official challenge system were as followings. Visual grounding was 0.8333, binary path validity was 0.0000, edge F1 was 0.5445, MESS was 0.3944, report score was 0.3966, background rejection was 0.8333, input sensitivity was 0.8333, cross region consistency was 0.8333, and overall score was 0.5444. The grounding modules showed nice performance, but the performance of the reasoning module was poor and we have to improve the performance.

5 Conclusions

In this paper, in accordance with the MICCAI 2026 REG² Challenge, we proposed a two-module framework that mimics the thought process of a pathologist. The reasoning module used a minimum path coverage algorithm to extract and learned optimal diagnostic pathways from a DAG. The grounding module performed region-of-interest (ROI) identification that was robust to image perturbations through binary classification using consistency loss. The overall score for Test Phase 1 in the official challenge system was 0.5444.

Although the grounding performance is relatively high, the quality of the reports and the MESS scores remain at a low level. Further analysis in future research is needed to determine whether this gap stems from the limitations of LLMs, a lack of alignment between images and language, or information loss resulting from the compression of visual tokens.

References

1. https://reg2026.grand-challenge.org/
2. https://zenodo.org/records/19848983/
3. Chen, R.J., et al.: Towards a General-Purpose Foundation Model for Computational Pathology. Nature Medicine 30(4), 1017–1026 (2024)
4. Lu, M.Y., et al.: Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2023, pp. 19764–19775 (2023)
5. Von Bergen, G., et al.: Med-CoT: Towards Proactive Clinical Agentic LLM. In: Proceedings of the 2025 IEEE International Conference on Big Data (BigData), pp. 8225–8229. IEEE, Macau, China (2025)
6. Zhu, C., et al.: MedEyes: Learning Dynamic Visual Focus for Medical Progressive Diagnosis. In: Proceedings of the Fortieth AAAI Conference on Artificial Intelligence (AAAI-26). AAAI Press (2026)
7. Salcan, et al.: Scalable Training of Spatially Grounded 2D Vision-Language Models for Radiology. arXiv preprint arXiv:2606.20477 (2026)
8. Liu, G., et al.: How Do Medical MLLMs Fail? A Study on Visual Grounding in Medical Images. arXiv preprint arXiv:2603.14323 (2026)
9. Lu, M.Y., et al.: A visual-language foundation model for computational pathology. Nature Medicine 30(3), 863–874 (2024)
10. https://qwen.ai/blog?id=qwen3.5

11. Huang, Z., et a.: A visual-language foundation model for pathology image analysis using medical Twitter. Nature Medicine 29(9), 2307–2316 (2023)
12. Yang, A., et al.: Qwen2.5 Technical Report. arXiv preprint arXiv:2412.15115 (2024)