Graph-Constrained Visual Grounding for Pathology Reasoning

Sujatha Kotte, Animesh Sanjay Sinha, Dhandapani Nandagopal, Vidushi Walia, Megha Sivakumar, Pranesh Kulasekhar, Vangala G. Saipradeep ⋆, and Naveen Sivadasan

TCS Research, Hyderabad, India
{kotte.sujatha,animeshs.sinha,n.dhandapani,vidushi.walia,megha.sivakumar, pranesh.kulasekhar,saipradeep.v,naveen.sivadasan}@tcs.com

Abstract. Pathology report generation from whole-slide images (WSIs) requires both accurate evidence selection and faithful diagnostic reasoning. We propose Graph-Constrained Visual Grounding for Pathology Reasoning, a compact framework for the REG2026 reasoning and grounding tasks. Annotated chain-of-thought questions are canonicalized into a directed diagnostic graph, where graph-constrained routing guides intermediate question answering and report generation. Frozen CONCH tile embeddings, TITAN slide embeddings, and PubMedBioBERT question embeddings are integrated through question-conditioned attention for evidence retrieval, while a LoRA-adapted BioGPT decoder generates answers and reports. For ROI-level grounding, the same model is reused without additional training as a single-step 5× ROI VQA system with sufficiency-aware prompting. On the held-out validation split, the framework achieved a workflow reasoning score of 0.759. On Challenge Test Phases 1 and 2, it achieved workflow reasoning scores of 0.620 and 0.538, and visual grounding scores of 0.892 and 0.850, respectively, indicating strong alignment between generated responses and visual evidence while limiting unsupported diagnostic claims. These results demonstrate the potential of graph-constrained routing for structured pathology reasoning, warranting further evaluation across broader clinical settings.

Keywords: Computational pathology · Whole-slide image · Visual grounding · REG2025 · REG2026 · Pathology Report generation · Pathology VQA · Diagnostic reasoning

1 Introduction

Pathology report generation from gigapixel whole-slide images (WSIs) extends beyond a conventional image-to-text generation task. In routine clinical practice, pathologists follow a structured diagnostic workflow that involves interpreting specimen context, verifying the presence of diagnostically relevant tissue, assessing morphological characteristics, applying disease-specific decision criteria, and ultimately synthesizing these observations into a pathology report. Consequently, generating a report directly from a WSI may yield linguistically plausible outputs while providing limited evidence that the reported findings originate from diagnostically relevant image regions.

⋆ Corresponding author: saipradeep.v@tcs.com

To address this limitation, REG2026[1] [2] evaluates pathology diagnostic reasoning models through two complementary dimensions of reasoning and grounding. REG2026 [1] [2] extends the earlier REG2025 [3] pathology report generation challenge by incorporating explicit pathologist reasoning traces and visual grounding evaluation in addition to final report generation. Metric A in the REG2026 Challenge evaluates slide-level workflow reasoning and final report generation, whereas Metric B evaluates whether a model can correctly answer a question from a local 5× ROI while avoiding diagnostic claims derived from background regions, masked areas, or insufficient tissue. We therefore separate the problem into two task-specific branches. Metric A is formulated as graph-constrained WSI reasoning over canonical questions. Metric B is formulated as sufficiency-aware ROI visual question answering (VQA). Both branches share question-conditioned visual grounding, but they differ in task structure: the former is a multi-step workflow graph and the latter is a single-step local image question.

The key contributions of this work are threefold: (i) a graph-constrained framework for modeling canonical diagnostic workflows in computational pathology; (ii) a question-conditioned visual evidence retrieval mechanism operating over pre-computed pathology representations; and (iii) a Bio-LLM-based answer generation module that integrates retrieved visual evidence to produce diagnostically grounded responses.

2 Materials & Methods

2.1 Overview

All experiments are conducted on the REG2026[1] [2] dataset. Trained models are evaluated on Metric A and Metric B of REG2026 Challenge. Figure 1 illustrates the architecture of our proposed Graph-Constrained Visual Grounding Pathology Reasoning model. The proposed model consists of: (i) Canonical Graph Construction (ii) Visual Feature Extraction - Frozen pathology foundation-model patch embeddings from CONCH [4] and slide embeddings from TITAN[5] (iii) Question Encoding - Biomedical question encoder using PubmedBioBERT [6] (iv) Question-Conditioned Visual Grounding (v) Fusion embedding. (vi) Graph-Constrained Workflow Reasoning - For next-questions prediction, and (vii) LoRA-adapted [7] biomedical LLM - BioGPT [8] for answer/report generation.

2.2 Data

The REG2026 challenge dataset consists of paired hematoxylin and eosin (H&E) whole-slide images (WSIs) and pathology reasoning traces represented as chain-of-thought (CoT) sequences. The WSIs were collected from multiple institutions across Korea, Japan, India, and Germany and were anonymized before release. Images were converted to TIFF format and standardized to a maximum magnification of 20×. Each case is accompanied by a structured pathology report and an associated reasoning trajectory describing the sequence of diagnostic questions and answers used to reach the final diagnosis. The dataset covers multiple organ systems, including breast, colon, lung, prostate, stomach, urinary bladder, and uterus, with both neoplastic and non-neoplastic diagnostic categories.

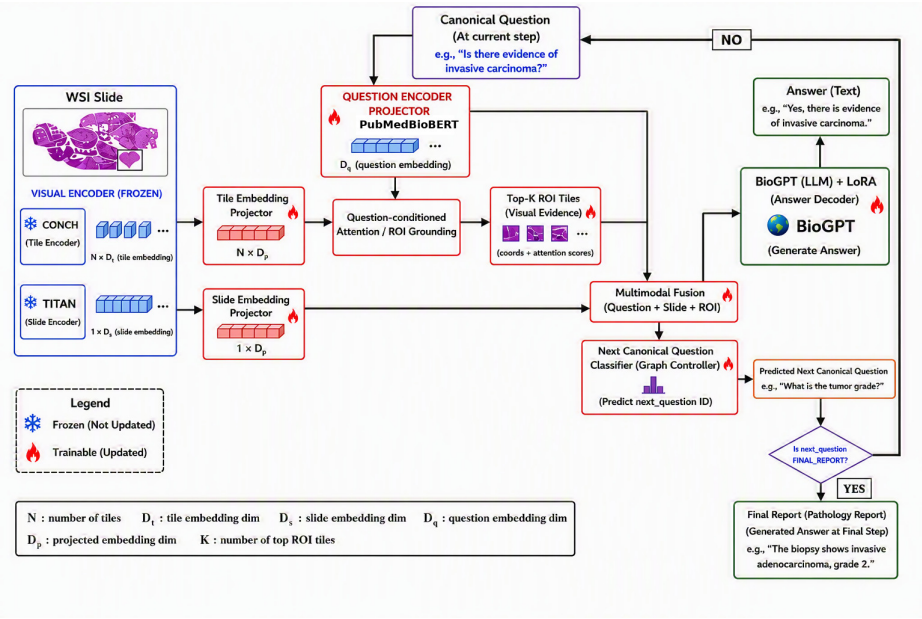

Fig. 1. Architectural flow of the proposed Graph-Constrained Visual Grounding Pathology Reasoning model.

The complete dataset contains 11,647 cases and is divided into a training set, debug and two independent test phases - test phase1 and test phase2. For Metric A, the training set consists of 11220 samples, each training sample includes a WSI, CoT - a sequence of canonical diagnostic questions, corresponding answers, observed reasoning transitions, and a final pathology report. The training split was used for model development, while evaluation follows the official REG2026 protocol using challenge-specific reasoning, grounding, and report-generation metrics.

During preprocessing, 1,050 WSIs were excluded due to slide-level input issues(e.g., missing metadata, unreadable files, or workflow incompatibilities). The remaining 10,170 annotated WSIs were stratified into training (n = 8,135) and internal validation (n = 2,035) sets using an 80:20 split. Exclusion rates varied across organs and were highest for prostate cases. To improve robustness, the inference pipeline handles WSI loading and preprocessing failures by returning an empty chain-of-thought ([]) instead of terminating execution. Slides with missing objective-power metadata were retained by assuming a 20× magnification. Organ-wise data distributions are provided in Supplementary Figure. 3.

2.3 Canonical Graph Construction

To model diagnostic reasoning as a constrained decision process, we constructed a single global directed Canonical Graph from the annotated chains of thought (CoTs)

available for all 11,220 challenge cases whike model training and validation were performed on the 10,170 WSIs that remained after preprocessing exclusions. The graph follows the official challenge protocol by retaining the provided canonical question and next_question strings whenever these strings are evaluated. Therefore, workflow questions are not semantically merged or clustered using embeddings, and questions with different canonical strings remain separate nodes even if they appear conceptually similar. Each distinct canonical question defines a node, and each observed ordered pair of consecutive questions defines a directed edge. Repeated transitions across different cases are consolidated into one edge, whereas multiple next questions observed after the same current question are all retained as permissible outgoing successors.

In this way, the graph captures the branching diagnostic pathways present in the annotated CoTs while preserving exact compatibility with the challenge annotations. To support multiple valid entry points, canonical root questions observed in the workflows, such as "What is the organ?" and "What is the procedure?", are treated as allowed starting nodes, and the graph terminates at the final-report stage. The graph aggregates CoTs across all organs rather than building separate organ-specific graphs; consequently, shared reasoning steps can be reused across organs, while organ and diagnosis specific workflows appear as overlapping subgraphs within the same structure. During inference, the graph does not impose a fixed diagnostic path. Instead, at each step it restricts candidate next questions to successors observed after the current node, and the WSI-derived visual context guides the model in selecting the appropriate transition and generating the corresponding answer. This design provides a unified yet graph-constrained reasoning workflow that preserves organ-specific diagnostic pathways and prevents unsupported transitions not observed in the annotated pathology reasoning sequences.

Table 1. Performance of the model in various phases.

	No. of Slides	Metric A	Metric A Sub-Scores				Metric B	Metric B Sub-Scores			Overall Score
			Binary Path Validation	Edge F1	MESS	Report Score		Background Rejection	Input Sensitivity	Cross Region Consistency	
Validation	2035	0.7587	0.25	0.7951	0.7211	0.8184	-	-	-	-	-
Debug	7	0.7947	0.1429	0.8291	0.7280	0.8921	0.9083	1.0000	0.9167	0.8333	0.8288
Test Phase 1	350	0.6202	0.1286	0.7634	0.6405	0.5616	0.8917	1.0000	0.7500	0.9167	0.7017
Test Phase 2	70	0.5383	0.1143	0.7584	0.5655	0.4110	0.8500	1.0000	0.8333	0.7500	0.6323

2.4 Visual Feature Extraction

WSIs were analysed using the Trident [9] pipeline. WSIs are processed using a tissue-aware patch extraction pipeline. Non-tissue regions are removed, 512×512 patches are extracted near 20× magnification, and the patches are encoded with a frozen CONCH pathology foundation model. For efficiency, visual features are precomputed before training. Each WSI is represented by a set of tile embeddings

$$V = \{v_i\}_{i=1}^{N}, \quad v_i \in \mathbb{R}^{d_v}, \tag{1}$$

where the current configuration uses $d_v = 768$. A global slide representation $s \in \mathbb{R}^{768}$ is extracted using the TITAN [5] slide encoder to capture overall tissue context.

The tile features and slide context are projected before the attention block, not inside the attention-section formulation. Specifically, tile vectors are mapped as

$$h_i^v = \mathrm{GELU}(\mathrm{LN}(W_v v_i)), \tag{2}$$

and the slide vector is mapped as

$$h^s = \mathrm{GELU}(\mathrm{LN}(W_s s)). \tag{3}$$

The hidden dimension is 512. At inference time, memory-efficient grid sampling with background filtering retains only informative tissue coordinates.

2.5 Question encoding and latent projection

Each canonical diagnostic question is encoded using a biomedical language encoder, PubMedBioBERT. Given a question q, the corresponding semantic representation is

$$e_q = f_q(q), \tag{4}$$

where $f_q(\cdot)$ denotes the biomedical question encoder. The question representation is down-projected into the same latent space as the tile features:

$$h^q = W_q e_q. \tag{5}$$

This projected question vector provides the task-specific context used by the visual grounding module and downstream reasoning heads.

2.6 Question-Conditioned Visual Grounding

Diagnostic evidence varies depending on the current reasoning step. Therefore, attention is conditioned explicitly on the active question. In the implemented model, attention is deliberately simple: it does not concatenate slide context into the tile scoring function. Instead, each projected tile feature is combined with the projected question vector using tanh attention:

$$r_i = \tanh(h_i^v + h^q), \qquad \ell_i = w_a^\top r_i, \tag{6}$$

where w_a is a learned attention scoring vector. Tile attention weights are then computed by

$$\alpha_i = \frac{\exp(\ell_i)}{\sum_{j=1}^{N} \exp(\ell_j)}. \tag{7}$$

The question-specific pooled visual representation is computed from the top-K attended tiles, where K=40

$$z_q = \sum_{i=1}^{K} \alpha_i h_i^v. \tag{8}$$

Finally, the pooled visual evidence, projected slide context, and projected question vector are fused:

$$h_t = \mathrm{MLP}([z_q, h^s, h^q]). \tag{9}$$

The top-k, where $k = 40$, tiles by attention weight are retained with indices, scores, and WSI coordinates for grounding and visual prefix construction. In the remainder of this manuscript, the top-k selection strategy with $k = 40$ is referred to as top-40.

2.7 Graph-Constrained Workflow Reasoning

For REG2026 Challenge Metric A, canonical reasoning traces form a directed graph $G = (Q, E)$ of diagnostic questions and valid transitions. The next-question classifier operates on the fused embedding h_t from Eq. 9, combining attention-pooled tile evidence, projected slide context, and projected question embedding; the top-40 attended tiles are retained as the interpretable evidence set. At step t, the model predicts answer $\hat{a}_t$ and selects the next set of valid questions:

$$\hat{q}_{t+1} = \arg \max_{q' \in \mathcal{N}(q_t)} p(q' \mid h_t), \tag{10}$$

where $\mathcal{N}(q_t)$ denotes valid successors. The same fused embedding feeds the next-question and edge heads; invalid candidates are masked before loss computation during training and inference-time softmax. Validation uses threshold 0.2 and at most two next questions per step. Unlike unconstrained sequence generation, graph-based routing prevents invalid diagnostic transitions and ensures consistency with canonical pathology workflows. For branching nodes, multiple successors may be emitted when their confidence scores exceed a predefined threshold.

2.8 LoRA-adapted biomedical LLM

Edge validity and next-question prediction are handled by classification heads, whereas intermediate answers and final reports are generated by the BioGPT-LoRA decoder using visual-prefix conditioning. [7]. Only a small set of trainable adapter parameters is optimized while the backbone language model remains frozen.

Trainable parameters include projection layers, question-conditioned ROI attention block, next set of questions heads, and LoRA adapters, while the foundation visual encoders, and BioGPT [8] base model remain frozen throughout training. The model is trained for 60 epochs with the best checkpoint saved based on validation loss. The total loss is a combination of the edge loss, answer loss, and next-question cross-entropy loss.

2.9 Training strategy

The model is trained over CoT steps with answer generation, edge prediction, and next-question classification losses:

$$\mathcal{L} = \lambda_{ans}\mathcal{L}_{ans} + \lambda_{edge}\mathcal{L}_{edge} + \lambda_{next}\mathcal{L}_{next}, \tag{11}$$

Here $\mathcal{L}_{ans}$ is the BioGPT visual-prefix answer loss, $\mathcal{L}_{edge}$ is masked edge classification, and $\mathcal{L}_{next}$ is masked next-question cross-entropy. Gradients back-propagate through tile/slide/question projections, tile-attention, fusion MLP, next and edge heads, visual-to-LLM projection, and LoRA adapters. Visual encoders and the BioGPT [8] backbone remain frozen. The configuration uses training of 60 epochs, batch size 32 CoT steps, patience parameter of 20 epochs for saving the best checkpoint, learning rates 2×10^{-5} and 1×10^{-4} for graph modules. Training employed bf16 precision, 4,096 WSI tiles and 8 visual tokens. It uses the top 40 retrieved patches per canonical question as grounding evidence

2.10 Inference for Metric A and Metric B adaptation

For Metric A, inference is a masked graph rollout from canonical START questions. Each step recomputes question-conditioned attention on tile embeddings, saves top-40 evidence, generates an answer with BioGPT-LoRA visual prefixes, and selects next questions only from valid graph neighbors. Thresholded fallback permits at most two branches and 32 steps before exporting CoT, edges, evidence, and report. For Metric B, the same checkpoint is reused as single-step ROI VQA: graph rollout is disabled, each 5× ROI is the complete visual input, and ROI-status prompting returns answers while avoiding diagnostic answers for background or insufficient tissue.

3 Results

3.1 Evaluation Metrics

We have evaluated our approach following the REG2026 [1] Challenge evaluation protocol, which comprises two complementary assessment tracks: Workflow Reasoning (Metric A) and Visual Grounding (Metric B).

Table 1 presents the performance of the approach on internal validation set, Debug Phase, test Phase1 on Metric A (Workflow Reasoning) and on the Metric B (Visual Grounding). Corresponding distributions of the Metric A sub-scores are provided in Supplementary Figure 1 and 2. Since annotations required for evaluating Metric B (Visual Grounding) were not available during the validation phase, Metric B performance was not assessed on the validation dataset. We have included the predicted CoTs for 2 cases in supplementary Table S2 and S3.

3.2 Ablations

Supplementary Table S1 presents inference-time ablations of the visual-grounding and graph-routing components. Replacing the attention-ranked top-40 tiles with 40 randomly selected tiles reduced Metric A from 0.759 to 0.520 and the finalreport score from 0.818 to 0.485. Retaining the attention-ranked top-40 tiles but assigning uniform weights reduced Metric A only slightly, from 0.759 to 0.755. These findings indicate that question-conditioned tile ranking provides the main performance benefit, while attention-weighted aggregation within the selected subset provides a smaller additional improvement. Disabling graph masking reduced Metric A to 0.428, Edge F1 to 0.509, and the final-report score to 0.265, demonstrating the importance of graph-constrained decoding for maintaining valid workflow progression.

Following the challenge protocol, the canonical graph was initially constructed using all 11,220 cases, as the task involves a closed set of predefined question pairs shared across the dataset. To evaluate potential information leakage, we additionally constructed the graph using only the 8,135 training slides. The ablation study on canonical graph construction shows that the training-only graph achieved a score of 0.80 , compared to 0.76 with the graph built from all cases. This result indicates that the proposed method does not benefit from information in the validation or test sets, suggesting that the canonical graph primarily captures the predefined question relationships rather than introducing information leakage.

4 Discussion and Conclusion

The proposed framework demonstrates the feasibility of reasoning-enabled pathology report generation by combining graph-constrained diagnostic workflows, question-conditioned visual grounding, and biomedical language models. Results show that the model can generate reasoning-consistent intermediate answers and coherent reports, while the ablation studies highlight the importance of graph-constrained reasoning and evidence-guided tile selection for workflow consistency and report quality. Furthermore, the performance obtained using a graph constructed only from training data suggests that the canonical graph captures predefined diagnostic relationships rather than introducing information leakage.

The performance gap between internal validation and Test Phases 1 and 2 likely reflects differences in dataset composition and case difficulty. Organ-wise analysis of the internal validation set (Supplementary Table S4) revealed substantial variability, with particularly low BPV scores in breast, cervix, lung, and bladder tissues, which may partly explain the lower overall BPV performance observed across both validation and test cohorts. A more comprehensive analysis will be possible once detailed test-set composition becomes available.

Future work will focus on evaluating larger biomedical and pathology-specific language models, exploring multiscale image representations, and improving reasoning robustness. We also plan to strengthen ROI-level validation, better capture rare reasoning paths, and incorporate challenging negative samples. These enhancements may further improve reasoning quality and visual grounding in generated pathology reports.

Reproducibility: To ensure reproducibility, the complete training code, model weights, and inference scripts are available in our private GitHub repository https://github.com/sujathakotte/REG2026_TPATH

Ethics Statement: This work was performed within the framework of the REG 2026 Challenge and exclusively used the challenge-provided de-identified datasets.

References

1. REG2026 Challenge Organizers: Pathologist Reasoning-Guided Report Generation Challenge (2026).
   Links: Grand Challenge: REG2026
2. Pathologist Reasoning-Guided Report Generation Challenge. Zenodo. International Conference on Medical Image Computing and Computer Assisted Intervention 2026.
   Links: Grand Challenge : REG2026 Zenodo
3. REG2025 Challenge Organizers: REport Generation in Pathology Using Pan-Asia Giga-pixel WSIs (2025).
   Links: Grand Challenge: REG2025
4. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., Parwani, A.V., Zhang, A., Mahmood, F.: A Visual-Language Foundation Model for Computational Pathology. *Nature Medicine* 30, 863–874 (2024). https://doi.org/10.1038/s41591-024-02856-4
   Links: Hugging Face: MahmoodLab/CONCH · GitHub: mahmoodlab/CONCH

5. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., Williamson, D.F.K., Chen, B., Almagro-Perez, C., Doucet, P., Sahai, S., Chen, C., Komura, D., Ishikawa, S., Le, L.P., Mahmood, F.: A Multimodal Whole-Slide Foundation Model for Pathology. *Nature Medicine* 31, 3749–3761 (2025). https://doi.org/10.1038/s41591-025-03982-3
Links: Hugging Face: MahmoodLab/TITAN · GitHub: mahmoodlab/TITAN
6. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare* 3(1), 1–23 (2021). https://doi.org/10.1145/3458754
Links: Hugging Face: microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In: *International Conference on Learning Representations* (2022).
Links: GitHub: microsoft/LoRA
8. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.-Y.: BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Briefings in Bioinformatics* 23(6), bbac409 (2022). https://doi.org/10.1093/bib/bbac409
Links: Hugging Face: microsoft/biogpt · GitHub: microsoft/BioGPT
9. Zhang, A., Jaume, G., Vaidya, A., Ding, T., Mahmood, F.: Accelerating Data Processing and Benchmarking of AI Models for Pathology. *arXiv preprint arXiv:2502.06750* (2025).
Links: GitHub: mahmoodlab/TRIDENT