Slide-Impression-Guided Modular Reasoning for Pathology Report Generation

Chia-Ping Chang[1], Yi-Chen Yeh[1], and Wen-Yih Liang[2]

[1] Taipei Veterans General Hospital, Taipei, Taiwan
cpchang6@vghtpe.gov.tw
[2] National Yang Ming Chiao Tung University, Taipei, Taiwan

Abstract. We describe our submission to the Pathologist REasoning-Guided REport Generation Challenge (REG^{2}2026), which asks an algorithm to turn a hematoxylin and eosin (H&E)-stained whole-slide image (WSI) into a pathology report together with the reasoning trajectory of question–answer pairs that leads to it, and to answer per-region visual-grounding questions. Rather than treating this as free-form generation, we adopt an impression-first, modular design that follows how a pathologist reads a slide: a frozen foundation model encodes each tile once, and attention-based multiple-instance-learning (MIL) heads first form a slide-level diagnostic impression. Conditioned on that impression, a visual-question-answering module answers the structured questions, proposes the clinically appropriate follow-up questions, and detects secondary findings. The trajectory is grown by a deterministic breadth-first traversal in which a learned reasoning module expands the diagnostic graph and answerer models answer each node. Out-of-distribution cases, flagged by the impression head's entropy, are handed to a slide-level vision–language model adapted with low-rank adaptation (LoRA) for a zero-shot organ and coarse impression. The report is then assembled by templating the established fields: organ, procedure, histological subtype, grade, and secondary findings. On in-distribution validation it attains a Workflow-Reasoning score of 0.9234; on the held-out Test Phase 1 leaderboard, under distribution shift, an overall 0.7652 (Workflow-Reasoning 0.7110, Visual-Grounding 0.8917), ranking ninth.

Keywords: Pathology report generation · Graph-of-thought · Computational pathology · Pathology foundation models.

1 Introduction

Computational pathology has been reshaped by foundation models developing along several parallel directions. Tile-level encoders [1, 5, 10, 6] provide general-purpose features for downstream tasks; vision-language models align images and diagnostic text in a shared space, at the patch level [9] and at the slide level [2]; and agentic systems such as CPathAgent [7] and SlideSeek [8] add autonomous multi-step reasoning for reporting. REG^{2}2026 targets this reasoning-guided setting, where the answer must be reached through a structured chain of reasoning

and written as a protocol-compliant report. The challenge scores not only the final report but the chain-of-thought (CoT) trajectory that produces it, and requires that trajectory to follow the question graph, with questions that fan out and back in and, for multi-diagnosis cases, are revisited under different parents.

Our central design decision is to be impression-first at the slide level. The information needed for almost every trajectory node is a slide-level result. The finer regional and histological details are then recognized by individual modules, each tailored to a type of question. The final report is rendered from a template rather than by open-ended generation, so that each trajectory node is produced by a dedicated module.

2 Methods

2.1 Dataset and metrics

The training set comprises about 11,220 WSIs across seven organs (Prostate, Breast, Colon, Stomach, Bladder, Lung, Cervix) from five countries and multiple scanners, each paired with a chain-of-thought Q&A annotation. The test phase has 350 cases (Test Phase 1) and 70 cases (Test Phase 2). Scoring combines a Workflow-Reasoning score (WFR) and a Visual-Grounding score (VG) as $\text{Overall} = 0.70 \cdot \text{WFR} + 0.30 \cdot \text{VG}$. The WFR weights four submetrics over the predicted reasoning graph, Binary Path Validity (BPV, 0.05), Edge-F1 (0.30), Mean Edge Semantic Similarity (MESS, 0.25), and a Final Report Score (0.40); the VG weights background rejection ($B1$), input sensitivity ($B2$), and tissue-versus-background separation ($B3$) as $0.30\,B1{+}0.30\,B2{+}0.40\,B3$. Full definitions are on the official challenge page.

Data use declaration. All whole-slide images and chain-of-thought annotations are REG^{2}2026 challenge data, de-identified by the organizers and used only under the challenge's data-use terms; no additional private data were used. The pretrained models we build on (H-Optimus-1, CONCH, TITAN, MUSK, MedCPT) are publicly released and used under their respective licenses.

2.2 Inference pipeline

Overview. Given a WSI at a single 20× magnification, we tile it into 224×224 patches and encode each patch once with a frozen tile foundation model, caching the resulting features. Every downstream module reads these cached features (Fig. 1). Because the diagnostic questions fan out into several follow-ups, fan back in, and are revisited under different parents, the reasoning trajectory is a directed acyclic graph (DAG) rather than a linear chain, and we refer to it as a graph of thought rather than a chain of thought. It is grown by breadth-first traversal: a reasoning module (M2) proposes the clinically appropriate follow-up question(s) at each node, and answerer models (M1) answer them, until the terminal "final report" question is reached. The report is then assembled from the answered nodes.

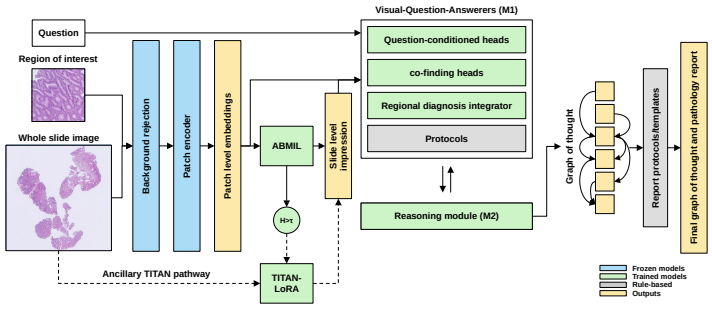

Fig. 1. Overview of the pipeline. A frozen patch encoder (H-Optimus-1) feeds an ABMIL slide-level impression, the M1 answerers and reasoning module M2 that grow the graph of thought, per-tile and region co-finding heads, and an entropy-gated ($H > \tau$) TITAN out-of-distribution path, from which report templates render the final report. Colour follows the legend. Abbreviations: ABMIL, attention-based multiple-instance learning; H, predictive entropy; τ, its threshold.

Tile encoding. Tiling and encoding use the TRIDENT toolkit [11]: a HEST tissue segmenter separates tissue from background, and tissue is patched at 20× into non-overlapping 224 × 224 tiles, each encoded by a frozen H-Optimus-1 [6] into a 1536-d feature. H-Optimus-1 is not fine-tuned; all task adaptation happens in the heads below. We also evaluated MUSK [9] for its vision-language alignment, but kept H-Optimus-1 as the stronger encoder for primary diagnosis. Features and tile coordinates are cached in HDF5 and re-used across all modules and both tasks, and the same segmentation rejects non-tissue background regions of interest (ROIs) at test time in the visual-grounding task.

Slide-level impression. An attention-based MIL aggregator (ABMIL, gated attention) pools the tile features into a slide embedding, and a linear classifier predicts the impression over a 139-class graded vocabulary from the training reports. Trained with single-label cross-entropy, it is deliberately a pure recognition head with no multi-label dilution, and its softmax entropy doubles as an out-of-distribution signal.

Visual-question answerers (Module 1, M1). The trajectory's structured questions (organ, procedure, histologic type or grade, presence gates) are answered from the same tile features by two complementary heads, because their answer spaces differ in shape. M1-cls is a closed-set classifier over each question's fixed legal answer set; it is the more accurate choice when the answer is one of a small, well-defined list, where an embedding-matching answerer tends to collapse onto closely related options. M1-vqa instead predicts an answer embedding, decoded by nearest neighbour over the full answer pool and conditioned on a small ordered context of anchor answers (organ and slide impression) plus the target question; it handles the large, open, or context-dependent answer spaces a fixed

classifier cannot, and it is the head reused for visual grounding. A per-question routing table sends each question to whichever head is more reliable.

Reasoning module (Module 2, M2). The diagnostic reasoning is a DAG with fan-out, fan-in, and revisits. M2 takes the answered history, with gate answers fed back, and predicts the set of appropriate next questions, its per-question fan-out capped to the maximum seen in training to avoid spurious branches. A breadth-first traversal expands the graph from the root, answering each node with the routed M1 head, until the terminal "What is the final pathology report?" node; revisits let the same question be re-asked under a new parent for multi-diagnosis workups.

Report generation. The final report is produced by filling structured templates with the organ, procedure, histologic type, and grade read from the answered trajectory, plus organ-specific derivations such as reconstructing the prostate Gleason score and grade group from the predominant and secondary patterns. Templating gives exact, reproducible field text and reduces the format drift of free-form decoding. It constrains wording and layout, not the prediction, so an incorrect field still yields an incorrect report.

Co-findings and multi-diagnosis. We address multi-diagnosis cases with two mechanisms, both on the frozen tile features. For focal or organ-specific findings we train small per-tile heads: a top-k-pooled multilayer perceptron for presence-type findings (microcalcification, granulomatous inflammation, organizing fibrosis) and for carcinoma in situ. Each fires at an organ- and primary-conditioned gate with a high-precision threshold, injecting the secondary diagnosis into the trajectory and report. The second mechanism is general and region-level: rather than target a hand-picked entity, it can surface an additional diagnosis from the base-diagnosis vocabulary seen in training, including one confined to a small area that whole-slide pooling would average away. Each $\sim$5 mm grid region is pooled by an ABMIL encoder into a region embedding, a per-base-diagnosis anchor prototype selected by the primary head cross-attends over the regions, and a head predicts the co-diagnosis set. Working at the region level keeps a locally-supported secondary diagnosis from being diluted by the dominant whole-slide signal, and the integrator emits only when its top prediction agrees with the primary, which keeps precision high.

Out-of-distribution management. The challenge's stated scope is broader than the released training data: its documentation names the uterus as an organ and diagnoses such as leiomyoma, whereas the training set covers only uterine cervix among uterine sites and contains no leiomyoma. Expecting out-of-distribution organs and diagnoses at test time, we add a hybrid path. When the primary head's predictive entropy exceeds a threshold, we invoke a TITAN slide-level vision–language model adapted with a low-rank adapter (LoRA), trained

on the slide transformer and contrastive projection head to lift zero-shot organ and coarse diagnosis on our distribution. Because the adapter is low-rank and sees only in-distribution organs, it leaves TITAN's open-vocabulary zero-shot ability [2] largely intact, which is what we rely on for unseen organs. This path supplies only a coarse organ and base diagnosis, while grade and detail come from the in-distribution heads.

Visual grounding. The visual-grounding task presents individual ROIs with questions on tissue presence, dominant content, and basic morphology, plus deliberately non-tissue background ROIs. We reuse the M1-vqa answerer on the ROI thumbnail features, but first pass each ROI through the same HEST segmenter used at tiling: if its tissue fraction falls below a threshold the ROI is returned a fixed non-diagnostic answer, and only tissue ROIs reach M1-vqa.

2.3 Building the models

Training data and ground-truth construction. All modules train on the challenge set under a deterministic, organ-stratified train/validation split, supervised from the chain-of-thought annotations. Every module reads the frozen H-Optimus-1 224×224 features, except the out-of-distribution adapter, which uses CONCH v1.5 [5] 512×512 features. The 139-class graded-diagnosis vocabulary is the distinct graded answers to the "$\#k$ diagnosis" questions, with a grade-stripped variant for the co-diagnosis heads. Answer and question texts are embedded once with the MedCPT query encoder [4] into fixed pools used as the answerers' cosine targets, the initial impression-classifier weights, and the nearest-neighbour table from a predicted embedding to a discrete answer. Each chain-of-thought is collapsed so repeated questions become one node, and a DAG attention mask lets a node attend to its ancestors but never to its own answer, preventing label leakage; question text is left uncanonicalised so training matches the raw-text edge scoring. A slide is positive for a co-finding entity only when it appears as a second-or-later diagnosis.

Training of each module. The impression head (ABMIL) trains with single-label cross-entropy on the graded #1 diagnosis, its classifier text-initialised from the answer embeddings. A base M1 answerer trains over the collapsed DAG with a cosine answer-embedding loss and an organ auxiliary loss, with scheduled sampling of its own predictions into the history, and its answers are decoded by nearest neighbour in the answer pool. M1-cls adds to this base a balanced per-question classification head over each closed question's legal answer set, which is more accurate there than the cosine match. M1-vqa instead fine-tunes the base with a reduced two-anchor context, the organ and slide impression, and a cosine loss on the target answer, its #1-diagnosis anchor occasionally replaced by a top-2 alternative for robustness. M2 trains on (node, ancestor-history) samples with a structural-weighted multilabel loss and two inference-mimicking relaxations: substituting M1-predicted answers and dropping DAG parent edges. The co-finding

Table 1. Frozen tile encoder for primary diagnosis: H-Optimus-1 vs. MUSK, over the 42 #1-diagnosis classes with ≥30 training slides (8439 train / 1028 validation slides).

	H-Optimus-1	MUSK
ABMIL top-1 accuracy	0.813	0.781
Linear probe (mean over organs)	0.686	0.624
Grade-of-neoplasm accuracy	0.784	0.796

and in-situ detectors are per-tile heads with positive-weighted cross-entropy calibrated to a high-precision threshold; the region integrator trains in two stages, a per-region ABMIL warm-started from the primary head then a cross-attention integrator over region embeddings against the case's base-diagnosis set. The TITAN adapter is a rank-16 LoRA [3] on the frozen encoder, trained with a CLIP-style loss against TITAN's own zero-shot text-prototype classifiers rather than a new closed-set head, so the reported open-vocabulary zero-shot ability is expected to be preserved.

Training platform and reproducibility. All heads consume cached H-Optimus-1 features (20×/224 px) and train in PyTorch on a single NVIDIA RTX PRO 6000 GPU. Inference is a fully deterministic, closed-container pipeline with no network access or external API. All code and reproduction steps are available in the repository linked in the supplementary material. Results are byte-reproducible under a fixed seed.

3 Results

3.1 Encoder and backbone comparison

We compared H-Optimus-1 and MUSK as the frozen tile encoder for the primary diagnosis (Table 1). The deployed impression head predicts a 139-class graded vocabulary spanning all diagnosis slots. For a controlled encoder comparison we used the #1-diagnosis answers and kept the 42 classes with at least 30 training slides, dropping a long tail too sparse to score an encoder fairly (8439 training and 1028 validation slides present in both encoders). With an identical ABMIL head, H-Optimus-1 reached 0.813 validation top-1 accuracy against 0.781 for MUSK, ahead on five organs and level on prostate and cervix. A head-independent logistic-regression probe on mean-pooled features gave the same ordering (0.686 vs. 0.624 averaged over organs). Grade-of-neoplasm accuracy was a near-tie (0.784 vs. 0.796), consistent with a shared visual ceiling for grading. H-Optimus-1's advantage is therefore on the primary-diagnosis identity that drives the whole cascade, and we kept it as the single frozen encoder.

Table 2. Workflow-Reasoning submetrics, WFR, and Visual-Grounding (VG) on validation, Test Phase 1, and Test Phase 2.

Split	BPV	Edge-F1	MESS	FinalRep	WFR	VG
Validation (all, n=1102)	0.759	0.961	0.931	0.911	0.9234	–
single-diagnosis (n=1033)	0.786	0.968	0.943	0.920	0.933	–
multi-diagnosis (n=69)	0.348	0.861	0.749	0.775	0.773	–
Test Phase 1 (overall 0.7652)	0.417	0.834	0.775	0.615	0.7110	0.8917
Test Phase 2 (overall 0.6948)	0.300	0.776	0.675	0.476	0.6069	0.9000

3.2 Replacing the whole primary with a slide-level model

We also tested whether a single slide-level vision–language model, TITAN [2], could replace the entire frozen-encoder-and-ABMIL primary at once, folding in-distribution recognition and out-of-distribution organ detection into one model. In distribution it did not match H-Optimus-1. Because TITAN is built on CONCH-v1.5 features, this comparison uses the validation subset for which those features were extracted. As a frozen slide backbone on a matched five-fold primary-diagnosis cross-validation over the 13 sufficiently-frequent classes (282 cases) it scored 0.837, behind H-Optimus-1 with ABMIL at 0.855, with CONCH and MUSK backbones at 0.830 and 0.801. Adapting TITAN with a rank-16 LoRA and a text-contrastive head on the 478 validation slides with CONCH v1.5 features raised its grade-stripped primary diagnosis to 0.815, still behind H-Optimus-1 at 0.848, and left the graded primary diagnosis further back at 0.720 against 0.802, the gap concentrated in the histological grade that the 512-pixel CONCH slide embedding resolves less sharply. Swapping the adapted TITAN in end-to-end on the same 478 slides for the organ and primary-diagnosis answers lowered validation WFR by 0.019, since primary-diagnosis errors propagate through the reasoning cascade. We therefore kept H-Optimus-1 for the in-distribution primary and use TITAN only in the entropy-gated out-of-distribution path.

3.3 Validation and official Test Phases 1 and 2

Table 2 reports the WFR and its submetrics on validation and the official Test Phase 1 and 2 leaderboards. Validation reaches WFR = 0.9234, with the remaining headroom concentrated in the multi-diagnosis workups (Table 2). Test Phase 1 scored an overall 0.7652 (WFR = 0.7110, VG = 0.8917), ranking ninth, and the final Test Phase 2 an overall 0.6948 (WFR = 0.6069, VG = 0.9000). All four workflow-reasoning submetrics decline under this cross-source shift, by roughly 0.19 to 0.46 from validation, while visual grounding stays near 0.90.

4 Discussion and Limitations

Our impression-first design replaces open-ended generation with dedicated modules and a report template, which suits a protocol-scored task. The same right-

tool-for-the-subtask principle, the core of our agentic design, runs through the pipeline: a per-question routing table answers each question with the head suited to its answer space, and the same idea governs how M2 grows the graph and how the co-finding and out-of-distribution modules are invoked. This is deterministic dispatch rather than an autonomously planned tool choice.

We kept H-Optimus-1 over MUSK, which we read as a task mismatch rather than a general weakness of MUSK. Our chain-of-thought is closed-answer and fully supervised, a multiple-instance classification problem in which MUSK's image-text alignment is redundant, and contrastive pretraining trades the fine discrimination needed to separate closely related diagnoses for coarser caption-level concepts. MUSK's zero-shot signal still helps for the non-epithelial and mesenchymal tumours our epithelial-biased head misses, but it does not beat H-Optimus-1 on the in-distribution primary diagnosis that drives the cascade.

Multi-diagnosis cases, where a secondary finding accompanies the primary, are clinically common and the hardest part of the metric, and are our largest remaining error source. A single missed secondary finding drops several trajectory edges and their answers at once, so its cost is out of proportion to its rarity. We attack them from two directions matched to how a finding presents: high-precision per-tile heads for focal, organ-specific findings that rarely fire on single-diagnosis cases, and a region-level integrator that recovers a more diffuse co-diagnosis without it being diluted by the dominant whole-slide signal.

Within visual grounding, the input-sensitivity term B2 is our weakest subcomponent. To keep the single backend the organizers require, we answer the grounding regions with the same M1-vqa model, which is applied here to single low-magnification region thumbnails that lie outside its whole-slide training distribution. B2 rewards invariance under appearance perturbation, so a model whose answers track the fine visual content of each region may receive a lower score on this term than one whose answers are more stable.

Because the test phases release no ground-truth answers, we cannot break the validation-to-test gap down by organ or failure mode, though its size already indicates clear room to improve cross-source generalization. Evaluating the entropy-gated out-of-distribution pathway on its own terms is harder still, as the graph-of-thought trajectory labels this challenge uses are uncommon in public pathology datasets, so building a comparable external benchmark would require substantial new annotation beyond this work.

5 Conclusion

We presented a modular, impression-first pipeline for reasoning-guided pathology reporting: a recognition head forms the slide-level impression, answerer and next-question modules grow and answer the reasoning graph, tile and region heads catch secondary findings, an entropy-gated TITAN path covers unfamiliar organs, and a template renders the report. Multi-diagnosis recall, visual-grounding stability, and out-of-distribution coverage remain the main limitations. For the last two, a single vision-language foundation model uniting H-

Optimus-1's patch-level detail with TITAN's slide-level breadth could let one answerer serve in-distribution reasoning, unfamiliar organs, and the grounding regions alike, which we see as a promising direction for future work.

References

1. Chen, R.J., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**, 850–862 (2024). https://doi.org/10.1038/s41591-024-02857-3
2. Ding, T., et al.: A multimodal whole-slide foundation model for pathology. Nature Medicine **31**(11), 3749–3761 (2025). https://doi.org/10.1038/s41591-025-03982-3
3. 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 (ICLR) (2022)
4. Jin, Q., Kim, W., Chen, Q., Comeau, D.C., Yeganova, L., Wilbur, W.J., Lu, Z.: MedCPT: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. Bioinformatics **39**(11), btad651 (2023). https://doi.org/10.1093/bioinformatics/btad651
5. Lu, M.Y., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024). https://doi.org/10.1038/s41591-024-02856-4
6. Scalbert, M., Saillard, C., Peeters, T., Gonzalez, L., Valter, D., Llinares-López, F., Mariet, Z.E., Jenatton, R.: H-optimus-1: A foundation model for computational histopathology. Cancer Research **86**(8_Suppl), LB174 (2026). https://doi.org/10.1158/1538-7445.AM2026-LB174
7. Sun, Y., et al.: CPathAgent: An agent-based foundation model for interpretable high-resolution pathology image analysis mimicking pathologists' diagnostic logic. arXiv preprint arXiv:2505.20510 (2025), neurIPS 2025
8. Weishaupt, L.L., Chen, C., Williamson, D.F.K., Chen, R.J., Jaume, G., Ding, T., Chen, B., Vaidya, A., Le, L.P., Lu, M.Y., Mahmood, F.: Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. arXiv preprint arXiv:2506.20964 (2025)
9. Xiang, J., et al.: A vision-language foundation model for precision oncology. Nature **638**(8051), 769–778 (2025). https://doi.org/10.1038/s41586-024-08378-w
10. Xu, H., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**, 181–188 (2024). https://doi.org/10.1038/s41586-024-07441-w
11. 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)