Diagnostic Pathology Reporting and Reasoning via Autoregressive and Retrieval-Based Chain-Of-Thought

Shubham Innani[1,2], Siddhesh Thakur[1,2], Michael Feldman[1,2], Spyridon Bakas[1,2,3,4,†]

[1] Division of Computational Pathology, Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA
[2] Indiana University Melvin and Bren Simon Comprehensive Cancer Center, USA
[3] Departments of Biostatistics and Health Data Science; Radiology and Imaging Sciences; Neurological Surgery, Indiana University School of Medicine, USA
[4] Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis, IN, USA
[†] Corresponding Author: spbakas@iu.edu

Abstract. Generating clinically useful pathology reports from gigapixel whole slide images (WSIs) requires computational models that identify diagnostically relevant evidence, follow a valid reasoning path, and produce meaningful reportable conclusions. The MICCAI 2026 REG^2 Challenge provides a comprehensive benchmark to facilitate innovation on this space. To this end, we developed two independent chain-of-thought (CoT) systems: i) Base-CoT and ii) R-CoT. Base-CoT is an autoregressive CoT model combining spatially contextualized CONCH patch embeddings, TITAN slide-level embeddings, and GECKO concept conditioning to generate structured diagnostic reasoning traces. R-CoT is a deterministic retrieval-based system decomposing reasoning generation into organ classification, diagnosis classification, and template retrieval conditioned on the predicted (organ, diagnosis) pair. R-CoT uses frozen UNI2-h patch features at 20× and 10× (capped at 512 tissue patches per scale), a linear head organ classifier on mean-pooled features, multi-scale per-organ gated-attention multiple-instance learning diagnosis ensembles, and a template bank mined from canonical training chains. Our quantitative performance evaluation for Base-CoT yielded a workflow ranking score of 0.445 on the official Test Phase 1 leaderboard, and 0.383 on our internal hold-out validation split, comprising 10% of the training data and stratified by organ with random sampling within each stratum. However, the R-CoT model on internal hold-out validation split yielded a workflow ranking score of 0.683. It achieved Test Phase 2 score of 0.6694 on the hidden test set, providing external validation of the retrieval-based approach.

Keywords: Computational pathology · Whole slide image analysis · Diagnostic reasoning · Chain-of-thought · Multiple instance learning · Report generation

1 Introduction

Pathology offers the final diagnostic confirmation in clinical care, and pathology reports are central to downstream treatment decisions. Producing such reports from H&E-stained whole slide images (WSIs) is challenging, as WSIs are gigapixel-scale images containing heterogeneous tissue regions, fine-grained morphologic features, and clinically important negative findings. A useful computational system must identify relevant evidence, preserve diagnostic logic, and avoid omissions that could potentially alter clinical interpretation. Recent advances in computational pathology have demonstrated the potential of artificial intelligence (AI) and deep learning methods to extract clinically relevant molecular, diagnostic, and prognostic information directly from routine H&E-stained histopathology slides [1–9].

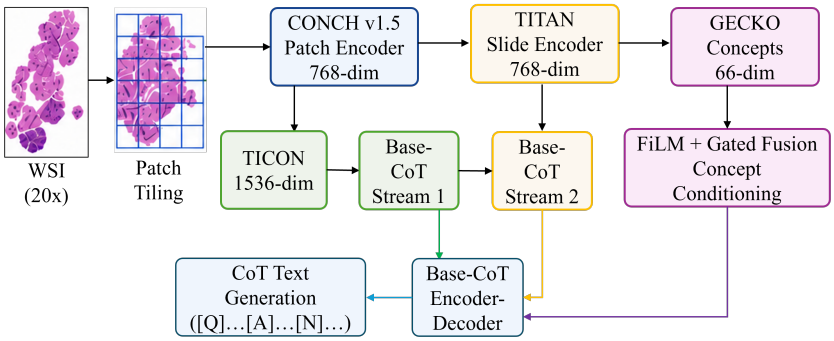
The MICCAI 2026 REG^2 Challenge [10] extends pathology report generation toward explicit diagnostic reasoning. Unlike earlier evaluations that emphasized lexical similarity [11, 12], REG^2 evaluates structured reasoning paths using measures of path validity, graph transition accuracy, semantic similarity, and final report quality. This formulation aligns with agentic and multi-step medical AI, where planning and evidence gathering, and clinical decisions are part of the task.

We developed two complementary systems, each conceived and implemented independently (Fig. 1). **Base-CoT** follows a generative strategy: it encodes WSI-derived patch, slide, and concept features and auto regressively emits a chain-of-thought (CoT) sequence over a closed CoT vocabulary and canonical question graph. **R-CoT** follows a discriminative/retrieval strategy: it predicts organ and diagnosis and retrieves a representative annotated CoT template for the predicted pair. The two systems were developed in parallel and are submitted together for comparative analysis. The retrieval design of R-CoT is motivated by the observation that training data already contains standardized, pathologist-style reasoning chains; reusing them preserves graph structure and reduces hallucination risk.

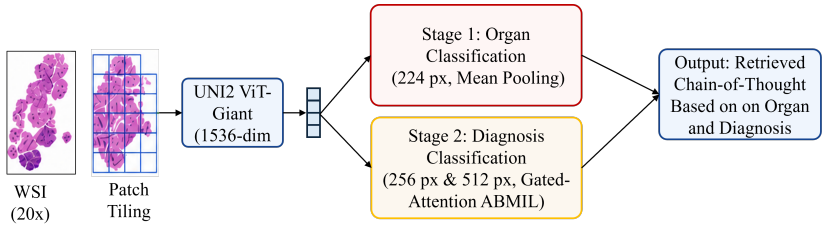
Our contributions are threefold. First, we present Base-CoT, a multimodal autoregressive baseline built on multi-level (patch-and slide-level) features leveraging CONCH, TICON, TITAN, and GECKO. Second, we present R-CoT, a deterministic classification-plus-retrieval system built on UNI2-h multi-scale ABMIL ensembles [13]. Third, we analyze the trade-off between generative flexibility and retrieval-based structural control for explicit WSI reasoning.

2 Related Work

Pathology foundation models have advanced rapidly. CONCH [11] and GigaPath [14] have demonstrated strong slide-level and patch-level representations for classification tasks. TITAN [12] extended these to multimodal slide-level embeddings suited for report generation. PRISM [15] combined patch features with autoregressive decoding for open-ended pathology report synthesis. These generative approaches excel at flexible text generation, but do not offer structural



(a) Base-CoT



(b) R-CoT

Fig. 1: **Two independent CoT systems. (a) Base-CoT** is an autoregressive reasoning pipeline built on CONCH v1.5 patch features, TICON-contextualized spatial features, TITAN slide embeddings, and GECKO concept features. The multimodal representation conditions a transformer encoder-decoder that generates CoT text. **(b) R-CoT** is a retrieval-based pipeline built on UNI2 ViT-Giant patch features. Organ classification uses mean-pooled 256-pixel features, followed by diagnosis classification with gated-attention ABMIL over 256- and 512-pixel features. The predicted organ-diagnosis pair retrieves a representative canonical CoT.

guarantees over reasoning graphs. UNI2-h [16] introduced a ViT-H encoder pretrained on large pathology datasets, providing strong discriminative features for downstream classification. Our R-CoT system builds on this discriminative strength, pairing UNI2-h features with multi-scale Attention Based Multiple Instance Learning (ABMIL) [17, 13] for organ and diagnosis prediction, and replaces generation with template retrieval to guarantee structural validity[13].

3 Challenge Task and Dataset

REG^2 provides H&E WSIs and paired CoT annotations derived from structured pathology reports. The dataset contains approximately 12,000 cases spanning seven organs: breast, colon, lung, prostate, stomach, urinary bladder, and

uterus. Diagnostic categories include malignant, pre-malignant, benign, and non-neoplastic entities. Slides were sourced from institutions in Korea, Turkey, Japan, India, and Germany, and reports follow the College of American Pathologists (CAP) protocol [18, 19] with diagnoses named per the WHO Classification of Tumours [20]. The official evaluation includes Test Phase 1 (350 slides) and Test Phase 2 (70 slides). We further constructed an internal hold-out validation set comprising 10% of the training data. Cases were stratified by organ class to preserve class balance, and within each organ stratum a random 10% subset was selected without replacement. These slides were excluded from both classifier training and template mining, ensuring no data leakage.

For each WSI, a system must output a structured reasoning chain and a final pathology report. Reasoning chains consist of canonical question-and-answer transitions, beginning with organ and procedure questions and ending in report generation. The workflow ranking score (Metric A) is defined per slide as

$$\text{Metric A} = 0.35 \times \text{BPV} + 0.35 \times \text{Edge-F1} + 0.30 \times \text{MESS}, \tag{1}$$

where BPV (Binary Path Validity) measures whether the predicted diagnostic path is structurally valid, Edge-F1 measures correctness of question-transition edges, and MESS (Morphological Embedding Semantic Similarity) measures semantic similarity of reasoning content. The challenge additionally reports a Final Report Score and a ROI-based visual grounding task (Metric B).

4 Methods

4.1 Base-CoT: Autoregressive Reasoning Graph Generation

Base-CoT generates CoT text from three feature streams. First, CONCH v1.5 patch embeddings [11] are extracted with TRIDENT [21] at 448px and contextualized by TICON (Tissue CONtext), which expands raw 768-dimensional patch features to 1,536-dimensional spatially aware features. Second, TITAN [12] provides a 768-dimensional slide-level embedding. Third, GECKO [22] operates on the raw CONCH features and produces 66-dimensional concept cosine similarities together with a 512-dimensional aggregated deep feature. The GECKO features condition the Base-CoT encoder through FiLM modulation and gated fusion. Fig. 1a summarizes the Base-CoT pipeline.

The fused streams are passed to a Base-CoT encoder-decoder that generates flattened CoT text in the format: $[Q]question[A]answer[N]next\_question$ A custom tokenizer restricts generation to a 455-token CoT vocabulary mined from training annotations. The canonical question vocabulary contains 92 questions, and inference clamps generated questions to this set to limit paraphrase drift and protect Edge-F1. The two required graph roots, "What is the organ?" and "What is the procedure?", are always emitted first.

4.2 R-CoT: Retrieval-Based Chain-of-Thought Reasoning

R-CoT is an independently developed system that avoids autoregressive language generation at inference. It decomposes the task into organ prediction, diagnosis

prediction, and retrieval of a canonical CoT template from the training set. Fig. 1b summarizes the R-CoT pipeline.

Feature extraction. WSIs are tissue-masked by thresholding and tiled into patch grids. A frozen UNI2-h ViT-H encoder [16] produces 1,536-dimensional patch embeddings. We extract patches at 256px for organ classification and at both 256px and 512px for diagnosis classification, capping at 512 tissue patches per scale (1,024 total) to satisfy the Grand Challenge 300-second inference constraint.

Organ classification. The 256px patch features are mean-pooled into a slide vector $\bar{\mathbf{z}} \in \mathbb{R}^{1536}$ and z-score normalized using statistics from the training data $(\boldsymbol{\mu}, \boldsymbol{\sigma})$. We train $N = 25$ independent linear heads on stratified cross-validation folds, each mapping $\mathbb{R}^{1,536}$ to the seven organ classes. At inference, logits are averaged:

$$\hat{y}_{\text{organ}} = \arg\max_c \frac{1}{N} \sum_{k=1}^{N} f_k \left(\frac{\bar{\mathbf{z}} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \right)_c . \tag{2}$$

Diagnosis classification. Diagnosis prediction uses per-organ gated-attention multiple instance learning (ABMIL) [17]. For a patch feature $\mathbf{z}_i$, we compute

$$\mathbf{h}_i = \text{ReLU}(\mathbf{W}_{\text{fc}} \mathbf{z}_i), \qquad \mathbf{h}_i \in \mathbb{R}^{768}. \tag{3}$$

Gated attention weights are

$$a_i = \text{softmax}(\mathbf{W}_{\text{attn}}(\tanh(\mathbf{V}\mathbf{h}_i) \odot \sigma(\mathbf{U}\mathbf{h}_i))) , \tag{4}$$

and the bag representation is $\mathbf{m} = \sum_i a_i \mathbf{h}_i$. A linear classifier maps $\mathbf{m}$ to the diagnosis classes for the predicted organ. Following a multi-scale MIL strategy [13], we train ABMIL ensembles at two scales for each of the seven organs. At 256px, we use 20-fold cross-validation (capturing fine-grained cellular morphology) and at 512px, we use 10-fold cross-validation (capturing broader tissue architecture). The 20/10 split reflects the greater discriminative information available at higher magnification while controlling the training time. This yields $7 \times (20 + 10) = 210$ models in total. At inference, the extracted feature tensors are passed through all folds at each scale, and a majority vote across all 30 predictions per organ yields $\hat{y}_{\text{dx}}$, where dx is diagnosis.

CoT template retrieval. A template bank is mined from the training CoT annotations. For each (organ, dx) pair, we select the representative chain whose length is closest to the median chain length for that pair. At inference, R-CoT retrieves with hierarchical fallback:

$$(\hat{y}_{\text{organ}}, \hat{y}_{\text{dx}}) \rightarrow \hat{y}_{\text{organ}} \text{ only} \rightarrow \text{global fallback}. \tag{5}$$

This guarantees a well-formed chain with valid question-transition edges by construction.

4.3 Visual Grounding for ROI Questions

For Metric B, we use MedGemma 1.5 4B IT [23] in a zero-shot setting. The ROI image and challenge-provided pathology question are passed to the model, and the generated answer is returned without task-specific fine-tuning. This branch is independent of the R-CoT template retrieval pipeline. ROI evidence does not currently influence template selection, which is a limitation addressed in the Discussion.

4.4 Implementation Details

Base-CoT was trained with contextualized CONCH/TICON patch features, TITAN slide embeddings, and GECKO concept and deep features. It uses a maximum CoT length of 512 tokens and validates every five epochs with a local scorer for BPV, Edge-F1, and MESS. Training ran on four A100 40GB GPUs with checkpoint-based resumption.

R-CoT was trained in four stages: UNI2-h feature extraction at 256px and 512px, 25-fold organ ensemble training, per-organ 30-fold (20+10) diagnosis ensemble training, and CoT template mining. The diagnosis ensemble contains 210 models across the seven organs (30 per organ). At inference, patch sampling is capped at 512 patches per scale (1,024 total); all 210 ABMIL models are preloaded at initialization, and inference completes in under 30s per slide on an A100 GPU.

5 Results

We evaluate our developed systems on internal held-out dataset (10% of the training data) with scores being computed with implementation of the official challenge metrics. Table 1 summarizes available scores. On the internal hold-out set, Base-CoT achieves a workflow ranking score of 0.383 (MESS=0.498, Edge-F1=0.507), while R-CoT reaches a workflow ranking score of 0.683 (MESS=0.761, Edge-F1=0.820). R-CoT outperforms Base-CoT on every metric, with the largest gains in MESS (+0.263) and Edge-F1 (+0.313). The high Edge-F1 follows directly from the retrieval design: each retrieved chain is a complete, human-authored workflow that preserves canonical question wording and transition structure by construction. MESS benefits similarly, as retrieved answers carry the vocabulary of training annotations rather than generated paraphrases. Base-CoT was evaluated on the official Test Phase 1 data through the challenge platform and obtained a workflow ranking score of 0.445. R-CoT was subsequently evaluated on the official hidden Test Phase 2 set and achieved workflow ranking score of 0.6694(Edge-F1=0.7645, MESS=0.6532). These results provide independent hidden-test validation of R-CoT. However, because Base-CoT and R-CoT were evaluated on different official test phases, their official scores should not be interpreted as a direct head-to-head comparison; the internal validation split (Table 1) remains the common evaluation set for comparison between the two systems.

Table 1: Workflow reasoning results. *Official*: challenge server scores on Test Phase 1 for Base-CoT and Test Phase 2 for R-CoT. *Internal Val*: our scorer on the held-out 10% split from training data. R-CoT Test Phase 1 was not submitted (Sec. 5). Ranking = 0.35×BPV +0.35×Edge-F1 +0.30×MESS. Higher is better.

Method	Split	MESS	Edge-F1	Report	BPV	Ranking
Base-CoT	Internal Val	0.498	0.507	0.200	0.160	0.383
Base-CoT	Test Ph. 1 (official)	0.571	0.655	0.239	0.191	0.445
R-CoT	Internal Val	0.761	0.820	0.564	0.420	0.683
R-CoT	Test Ph. 1	not submitted (see Sec. 5)				
R-CoT	Test Ph. 2	0.6532	0.7645	0.4270	0.3571	0.6694

On the internal validation set, 91.3% of cases were resolved at the exact (organ, dx) level, 6.8% fell back to organ-level templates, and 1.9% required the global fallback. Cases requiring organ-level fallback showed an average BPV reduction of 0.11 and a MESS reduction of 0.09 relative to the full-match cases, reflecting that organ-level templates carry the correct structural skeleton but lack diagnosis-specific answer content. Global fallback cases (rare diagnoses absent from training) showed larger degradation (BPV −0.19, MESS −0.14). From a clinical perspective, retrieval guarantees structural validity but does not guarantee content specificity. A template matched on (organ, diagnosis label) may not capture the morphologic nuance, grading, margin status, or secondary finding that determines clinical management for a given case. For example, two cases of invasive breast carcinoma NST Grade II may differ in lymphovascular invasion, HER2 status, or tumor size, yet receive the same template. This risk is inherent to the current median-chain retrieval strategy. Future work on similarity-based retrieval, selecting templates conditioned on morphologic feature proximity rather than diagnosis label alone, could substantially mitigate this limitation (see Section 6). The two systems occupy different points in the trade-off between flexibility and control. Base-CoT can generate novel reasoning chains conditioned on multimodal features, which may help when the target diagnosis is absent from the template bank. R-CoT is less flexible but deterministic, faster (under 30s/slide vs. token-by-token decoding), and free of graph invalidity for covered (organ, dx) pairs.

6 Discussion

We presented two independently developed systems to structured pathology reasoning for REG^2 2026: **Base-CoT**, which autoregressively generates diagnostic chains from multimodal WSI features, and **R-CoT**, which classifies organ and diagnosis and retrieves a canonical template. Base-CoT was built on CONCH, TITAN, and GECKO features, and R-CoT was built on UNI2-h multi-scale ABMIL classification. On our internal validation split, R-CoT substantially outperformed Base-CoT across all metrics (ranking 0.683 vs. 0.383), driven by the

structural guarantees of retrieval-based generation. Base-CoT obtained a Test Phase 1 ranking score of 0.445 on the challenge server, whereas R-CoT obtained a Test Phase 2 ranking score of 0.6694. R-CoT demonstrates that organ and diagnosis classification followed by canonical CoT retrieval yields strong Edge-F1 and MESS obtaining a score of 0.7645 and MESS of 0.6532 respectively.

Base-CoT's decoder operates over a closed 92-question vocabulary mined from training annotations. This restricts the system to the set of question types seen during training and does not support open-ended question formulations. However, for the REG[2] task the question graph is itself canonically defined, so this constraint is aligned with the evaluation protocol rather than being an architectural limitation. R-CoT's structural validity is guaranteed by construction, but content fidelity is bounded by the granularity of the template bank. Diagnoses within a category are rarely interchangeable in clinical practice, i.e., two cases with the same (organ, diagnosis) label may differ in grading, margin status, lymphovascular invasion, or secondary findings that determine patient management. The current median-chain retrieval strategy does not account for these intra-class morphologic differences. On our internal validation set, 8.7% of cases required organ-level or global fallback, with measurable drops in BPV (-0.11 to -0.19) and MESS (-0.09 to -0.14). Mitigating this risk would require feature-conditioned retrieval, selecting templates whose morphologic embedding is closest to the query slide, rather than matching on diagnosis label alone. This is identified as a priority for future work.

A remaining evaluation limitation is that Base-CoT and R-CoT were not evaluated on the same official hidden test phase. Base-CoT was evaluated on Test Phase 1, whereas R-CoT was evaluated on Test Phase 2. Consequently, differences between their official scores may reflect both methodological differences and variation between the two hidden test sets. Direct quantitative comparison between the systems therefore remains most appropriately based on the common internal hold-out validation set.

Template retrieval depends on training coverage. On the internal validation set, 8.7% of cases triggered organ-level or global fallback, with BPV reductions of 0.11-0.19 and MESS reductions of 0.09-0.14 relative to exact-match cases. Rare diagnoses absent from training always require fallback, and the resulting template may not capture the specific morphologic nuance or grading detail that determines case management. R-CoT also separates WSI-level reasoning from ROI question answering, and hence ROI evidence does not currently influence template selection. Base-CoT's question vocabulary is closed (92 questions), limiting flexibility to question types seen during training. Visual grounding (Metric B) is handled by a zero-shot MedGemma 1.5 4B that is entirely decoupled from the R-CoT classification and retrieval pipeline. This means that ROI-level evidence, the specific tissue region identified as diagnostically salient, does not influence template selection. Tighter integration, where attention maps or ROI embeddings condition template retrieval, is also considered future work.

Future directions to extend the presented systems is expected to prioritize: (1) morphologic similarity-based template retrieval to address intra-class content

variation; (2) ROI-conditioned chain refinement to integrate visual grounding evidence into template selection; and (3) same-phase evaluation of Base-CoT and R-CoT on a common hidden test cohort to enable controlled external comparison.

Acknowledgements Computational resources used in this research were partially supported by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

1. Baheti, B., Innani S., Nasrallah M.P., and Bakas S.: Prognostic stratification of glioblastoma patients by unsupervised clustering of morphology patterns on whole slide images furthering our disease understanding. Frontiers in Neuroscience 18 (2024): 1304191.
2. Baheti, B., Rai S., Innani S., Mehdiratta G., Bell W.R., Guntuku S.C., Nasrallah M.P., and Bakas S. Multimodal explainable artificial intelligence for prognostic stratification of patients with glioblastoma. Modern pathology 38, no. 9 (2025): 100797.
3. Innani, S., Bell, W.R., Nasrallah, M.P., Baheti, B., Bakas, S.: PATH-67. AI-Based WHO 2021 Classification of Adult-Type Diffuse Glioma Solely from H&E-Stained Slides. Neuro-Oncology **27**(Supplement_5), v256–v256 (2025)
4. Collins, K., Innani, S., Ebare, K., Saad, M., Siegmund, S.E., Williamson, S.R., Maclean, F., et al.: Artificial Intelligence–Based Classification of Renal Oncocytic Neoplasms: Advancing From a 2-Class Model of Renal Oncocytoma and Low-Grade Oncocytic Tumor to a 3-Class Model Including Chromophobe Renal Cell Carcinoma. Archives of Pathology & Laboratory Medicine **149**(10), 922–929 (2025)
5. Innani, S., Pitarch-Abaigar, C., Marwan, M.M., Harmsen, H., Bell, W.R., Makris, D., Bakas, S.: PATH-69. AI-Based Prognostic Risk Stratification of Adult-Type Diffuse Gliomas Using H&E-Stained Slides. Neuro-Oncology **27**(Supplement_5), v257–v257 (2025)
6. Innani, S., Bell, W.R., Nasrallah, M.P., Baheti, B., Bakas, S.: Artificial Intelligence Predicts 2021 WHO Glioma Subtypes from Whole Slide Images. Cancer Research **85**(8_Supplement_1), 6247–6247 (2025)
7. Innani, S., Bell, W.R., Harmsen, H., Nasrallah, M.P., Baheti, B., Bakas, S.: Interpretable Artificial Intelligence Based Determination of Glioma IDH Mutation Status Directly from Histology Slides. Neuro-Oncology Advances **7**(1), vdaf140 (2025)
8. Innani, S., Bell, W.R., Nasrallah, M.P., Baheti, B., Bakas, S.: AI-Driven WHO 2021 Classification of Gliomas Based Only on H&E-Stained Slides. Neuro-Oncology **28**(1), 282–296 (2026)
9. Innani, S., Baheti, B., Nasrallah, M.P., Bell, W.R., Bakas, S.: PATH-39. AI-Based Identification of Glioma IDH Mutational Status from H&E-Stained Whole Slide Images. Neuro-Oncology **26**(Suppl 8), viii187 (2024)
10. REG^2 2026 Challenge Organizers: Reporting and Evaluation of Generated Reasoning 2026. MICCAI Challenge (2026), `https://reg26.grand-challenge.org/`
11. Lu, M.Y., Chen, B., Williamson, D.F.K., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024)

12. Ding, J., Jiang, Z., et al.: A multimodal whole-slide foundation model for pathology (TITAN). Nature Medicine (2025). https://doi.org/10.1038/s41591-025-03982-3
13. Innani, S., Bell, W.R., Nasrallah, M.P., Baheti, B., Bakas, S.: Multi-scale whole slide image assessment improves deep learning based WHO 2021 glioma classification. In MICCAI Workshop on Computational Pathology with Multimodal Data (COMPAYL). 2024.
14. Xu, H., Usuyama, N., Bagga, J., 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
15. Shaikovski, G., Casson, A., Severson, K., et al.: PRISM: A multi-modal generative foundation model for slide-level histopathology. arXiv preprint arXiv:2405.10254 (2024)
16. Chen, R.J., Ding, T., Lu, M.Y., 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
17. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: ICML, PMLR, pp. 2127–2136 (2018)
18. College of American Pathologists: Cancer Protocol Templates. `https://www.cap.org/protocols-and-guidelines/cancer-reporting-tools/cancer-protocol-templates`, last accessed 2026/07/20
19. Torous, V.F., Simpson, R.W., Balani, J.P., et al.: College of American Pathologists Cancer Protocols: from optimizing cancer patient care to facilitating interoperable reporting and downstream data use. JCO Clin. Cancer Inform. **5**, 47–55 (2021). https://doi.org/10.1200/CCI.20.00104
20. WHO Classification of Tumours Editorial Board: WHO Classification of Tumours, 5th edn. International Agency for Research on Cancer, Lyon. `https://tumourclassification.iarc.who.int`, last accessed 2026/07/20
21. Kang, M., Song, A.H., et al.: TRIDENT: Efficient whole-slide image feature extraction and processing for computational pathology. Technical report (2024)
22. Filiot, A., et al.: GECKO: Generative encoder for contextualized knowledge in oncology. Technical report (2024)
23. Sellergren, A., et al.: MedGemma Technical Report. arXiv preprint arXiv:2507.05201 (2025)