# Reporting Without Generating: Retrieval-Based Chain-of-Thought Answering for Pathology

Lukas Buess and Andreas Maier

Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
lukas.buess@fau.de

**Abstract.** Reliable automated pathology reporting should yield not only an accurate final report but also a transparent, auditable trace of the diagnostic decisions behind it. We present RECAP (REtrieval-based Chain-of-thought Answering for Pathology), a training-free method that maps an H&E whole-slide image (WSI) to a retrieved reasoning graph consisting of canonical diagnostic question-answer steps that terminates in a structured pathology report. Each slide is encoded with the frozen TITAN foundation model, and the canonical reasoning graph is completed by $k$-nearest-neighbour ($k$-NN) retrieval over TITAN slide embeddings, so that the final report, case-variable answers, and reasoning-graph edges are all drawn from the training distribution. No generative language model is used during inference. Rather than assembling a prediction from independently retrieved components, a report-first stage commits to the single real case whose findings best match the draft and adopts its consensus reasoning chain, ensuring consistency between the retrieved reasoning graph and the final report. Because RECAP requires no task-specific training and encodes only the test slide at inference, it is inexpensive to deploy in resource-constrained settings and benefits directly from improvements to the underlying foundation models. On a held-out split of the released training data, RECAP attains a Metric A Reg2026-Score of $\approx$0.88.

**Keywords:** Chain-of-thought · Computational pathology · Foundation models · Retrieval-augmented generation · Whole-slide images.

## 1 Introduction

Diagnostic pathology reports are authored from gigapixel whole-slide images (WSIs); automating them could reduce reporting workload and standardise clinical language. Progress has been driven by pathology foundation models built on CLIP-style vision-language pretraining [9]: patch encoders such as CONCH [8], UNI [3], and H-Optimus [11], and slide-level models such as Prov-GigaPath [17], PRISM [13], and TITAN [5], which give strong representations without task-specific training. For report text, generative approaches range from slide-level models such as PRISM [13] to general medical large language models (e.g., Gemini for medicine [10], MedGemma [12], and OpenBioLLM [1]), which produce

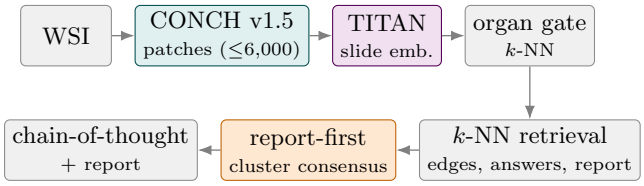


**Fig. 1.** RECAP pipeline (Metric A). A whole-slide image is tiled into CONCH v1.5 patches (capped at 6,000) and aggregated by TITAN into a slide-level embedding. An organ gate ($k$-NN) picks the organ; same-organ $k$-NN retrieves the graph edges, per-question answers, and report. Report-first then swaps this draft for the consensus chain-of-thought of the most similar report cluster.

free-text end-to-end. A complementary line of work is retrieval-based, grounding each prediction in real existing cases [2].

The REG2026 challenge frames automated pathology reporting as a reasoning task: a method must emit not only a final report but the intermediate question/answer steps a pathologist would follow, as a directed acyclic graph whose nodes are canonical questions and whose terminal node is the report [14]. Submissions are scored both on the exact reasoning-graph structure and on the content of the final report. The canonical question set is fixed and finite and the reference reports come from real cases, so the training annotations form a strong prior, motivating a retrieval-augmented [7] rather than free-text generative treatment of the reasoning graph [16].

We tackle this task with a training-free, retrieval-based method that uses no generative language model in the reporting path. Rather than assembling the trace from independently retrieved components, a report-first stage commits to a single real case and adopts its consensus chain, keeping the decision trace and the reported diagnosis mutually consistent. Our contributions are as follows:

- a strong training-free retrieval baseline for structured whole-slide pathology reporting, using frozen foundation model representations and no task-specific parameter optimisation or generative language model, which improves automatically as the underlying foundation models improve;
- an analysis showing that simple case retrieval accounts for most of the achieved performance, highlighting representation quality and accurate report retrieval as key levers for further improvement.

## 2 Methods

Figure 1 summarises RECAP. Let $x$ be a whole-slide image, with organ $o_i$, ground-truth chain-of-thought graph $G_i$, and report $r_i$ for each training slide $i$ in the labelled pool $\mathcal{D} = \{(x_i, o_i, G_i, r_i)\}_{i=1}^{N}$. RECAP produces a graph $\hat{G}$ for $x$ purely by nearest-neighbour retrieval against $\mathcal{D}$. In Sections 2.1-2.3 we focus on the workflow-reporting task (Metric A); the separate patch-level grounding task (Metric B) is discussed in Section 2.4.

### 2.1 Slide encoding and organ gate

A slide $x$ is tiled into patches, embedded by the frozen CONCH v1.5 encoder $\phi$, and aggregated by the frozen TITAN model $T$ into a slide embedding

$$\mathbf{z} = T\big(\phi(x)\big) \in \mathbb{R}^{768}. \tag{1}$$

Likewise, $\mathbf{z}_i = T(\phi(x_i))$ is precomputed for every slide $x_i$ in the retrieval pool $\mathcal{D}$. Similarity is measured by cosine similarity between $\ell_2$-normalised embeddings. The organ is predicted by a $k$-nearest-neighbour vote. Let $\mathcal{N}_{\text{org}}(\mathbf{z})$ denote the $k$ training slides most similar to the query embedding $\mathbf{z}$. RECAP predicts the most frequent organ label among these neighbours,

$$\hat{o} = \text{mode}\{\, o_i \mid i \in \mathcal{N}_{\text{org}}(\mathbf{z}) \,\}. \tag{2}$$

This image-to-image vote over the training embeddings is more reliable than a zero-shot text classifier (Section 4).

### 2.2 Draft chain-of-thought by retrieval

After predicting $\hat{o}$, retrieval is restricted to training slides from that organ. Let $\mathcal{N}_{\text{ret}}(\mathbf{z}, \hat{o})$ denote the $k$ most similar same-organ slides. RECAP traverses the fixed canonical question graph $\mathcal{G}_{\hat{o}}$ and constructs a draft reasoning graph $\hat{G}_0$ from these neighbours. **Edges:** an edge $e$ (a directed link from one canonical question to the next) is kept if it occurs in at least a fraction $\tau$ of the retrieved graphs,

$$\hat{E}_0 = \{\, e \in \mathcal{G}_{\hat{o}} : \text{freq}_{\mathcal{N}_{\text{ret}}}(e) \geq \tau \,\}, \tag{3}$$

where $\text{freq}_{\mathcal{N}_{\text{ret}}}(e)$ is the fraction of retrieved graphs containing $e$; the report node is always kept. **Answers:** each non-final question takes the most frequent answer among the retrieved neighbours that contain that question. **Report:** the draft report $\hat{r}_0$ is the most frequent report among the retrieved neighbours,

$$\hat{r}_0 = \text{mode}\{\, r_i \mid i \in \mathcal{N}_{\text{ret}}(\mathbf{z}, \hat{o}) \,\}. \tag{4}$$

### 2.3 Report-first prediction

The draft $\hat{G}_0$ can be internally inconsistent because its edges, answers, and report are determined independently. Report-first instead selects a coherent diagnostic pattern from the retrieval pool. Writing $F(G)$ for the findings of a reasoning graph (its non-final (question, answer) pairs), we rank the same-organ slides by their finding overlap with the draft,

$$s_i = \Big| F(\hat{G}_0) \cap F(G_i) \Big|. \tag{5}$$

Let $\mathcal{N}_{\text{match}}$ denote the $k$ highest-scoring same-organ slides according to $s_i$. The final report $r^\star$ is the most frequent report among these matches. RECAP then selects the most frequent complete reasoning graph among all retrieval-pool slides carrying $r^\star$. Because slides sharing a report have almost identical reasoning graphs in the data, the resulting graph represents one coherent diagnostic pattern rather than a graph stitched from independently retrieved components.

### 2.4 Patch-level visual grounding

The challenge's second axis (Metric B) scores short answers to region-of-interest (ROI) questions on three sub-metrics: background rejection (B1), input sensitivity under random patch masking (B2), and cross-region consistency (B3). We answer with two fixed templates, gated by the same near-white tissue check the slide-level reporting task (Metric A) uses to select patches: an ROI whose tissue fraction is below 0.05 gets the background template, every other ROI the foreground template.

## 3 Experimental setup

**Models.** We use the frozen TITAN slide encoder because the task requires slide-level predictions from the complete WSI. Its large-scale pathology pretraining provides transferable representations without task-specific optimisation. TITAN embeds 512×512 level-0 patches with CONCH v1.5 [8] (306.1M parameters; capped at 6,000 per slide by uniform subsampling) and aggregates them with its 48.5M-parameter slide encoder into a 768-dimensional embedding. The complete frozen WSI backbone comprises approximately 355M parameters.

**Data.** From the 11,220 slides, 11,189 yield a valid TITAN embedding; we hold out 500 (uniform random) for evaluation (498 of which embed successfully).

**Metric.** The workflow score (Metric A) is 0.05 BPV+0.30 Edge-F1+0.25 MESS+ 0.40 FinalReport, where Binary Path Validity (BPV) is 1 only on an exact edge-set match, Edge-F1 scores the (question, next-question) edges, MESS is the token Jaccard on non-final answers, and FinalReport blends ROUGE-L, BLEU-4, keyword overlap, and PubMedBERT [6] cosine similarity. Metric B (visual grounding) is scored on B1-B3 as defined in Section 2.4; the overall challenge score combines the two axes as 0.7 Metric A + 0.3 Metric B.

## 4 Results

We report results on two evaluation settings: a held-out split of the released training data and the official REG2026 hidden test sets.

**Organ gate.** The image-to-image $k$-NN vote classifies the organ almost perfectly on the held-out split (99.4%), well above a zero-shot TITAN text classifier (94.8%), confirming the design choice in Section 2.

**Held-out results.** Table 1 reports the full RECAP method on the 498-slide held-out split, scored by the official evaluator.

**Organ-level evaluation.** Figure 2 breaks the Metric A score down by organ. Performance is uniform across the seven organs (overall ranking 0.86-0.90; lung strongest, cervix weakest). Edge-F1 and MESS remain high and stable (0.89-0.96 and 0.89-0.95), indicating robust structural and answer retrieval. Binary Path Validity (0.65-0.85) and the final-report term (0.74-0.86) vary most; with the report carrying the largest weight (0.40), overall ranking follows it closely, leaving cervix weakest overall.

**Table 1.** The full RECAP method on the 498-slide held-out split of the training data (Metric A components and weighted contributions).

| Component | Score | Weight | Contribution |
|---|---|---|---|
| BPV (Binary Path Validity) | 0.7329 | 0.05 | 0.0366 |
| Edge-F1 | 0.9321 | 0.30 | 0.2796 |
| MESS (non-final answers) | 0.9274 | 0.25 | 0.2319 |
| Final report | 0.8245 | 0.40 | 0.3298 |
| **Metric A (ranking score)** | | | **0.8779** |

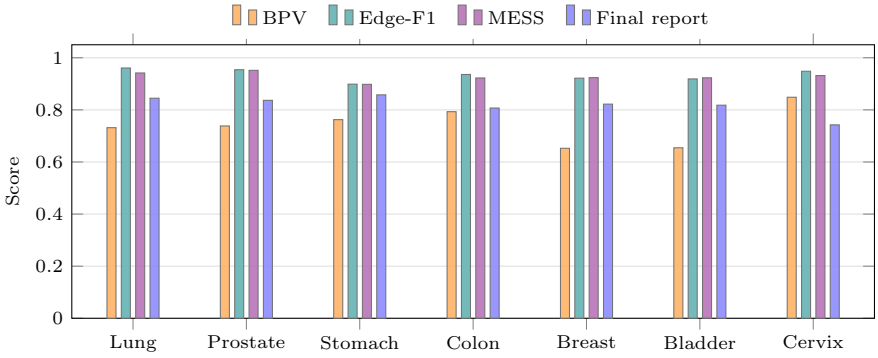


**Fig. 2.** Metric A components of the RECAP method on the 498-slide held-out split, by organ (organs ordered by overall ranking, 0.898 for lung down to 0.857 for cervix). Edge-F1 and MESS are high and stable across organs; the widest variation is in Binary Path Validity and the final-report term, the latter carrying the most ranking weight.

### 4.1 Ablation

Table 2 re-scores the held-out set for two kinds of comparison: whole-pipeline retrieval baselines that swap in a simpler strategy, and ablations that turn off one component of the full RECAP method.

**Retrieval baselines.** Plain retrieval is already most of the method. Copying the single closest slide's chain-of-thought verbatim (image-to-image naive 1-NN) scores 0.871, within 0.007 of the full method and on par with the per-question pipeline; most of the score comes from retrieving a coherent real case, not from how it is assembled. The choice of retrieval space does matter: matching the test slide against the training reports in TITAN's cross-modal space (image-to-text) drops the naive 1-NN to 0.697, so TITAN's slide embeddings discriminate cases far better than its slide-report alignment. The random-retrieval floor (0.305) confirms the task is non-trivial: the metric is far from saturated, confirming that strong performance requires accurate retrieval.

**Component ablations.** Against the baseline, edge voting is by far the dominant lever ($-0.153$): instead of keeping only the edges that occur in at least a fraction $\tau$ of the neighbours' graphs, disabling it emits every edge of the organ's

**Table 2.** Ablations and retrieval baselines on the 498 held-out slides. Component ablations remove one part of the full report-first method, while whole-pipeline baselines replace it with a simpler retrieval strategy. $\Delta$ denotes the drop from the full method.

| Configuration | Ranking | $\Delta$ vs full |
|---|---|---|
| RECAP (full method) | 0.8779 | — |
| *Component ablations* | | |
| − report-first (per-question draft) | 0.8697 | −0.0082 |
| − organ routing | 0.8450 | −0.0330 |
| − edge voting (structure) | 0.7251 | −0.1528 |
| *Whole-pipeline retrieval baselines* | | |
| Naive 1-NN (image–image) | 0.8707 | −0.0072 |
| Naive 1-NN (image–text) | 0.6973 | −0.1807 |
| Random retrieval | 0.3050 | −0.5729 |

canonical graph, so the predicted graph is never an exact match; BPV collapses ($0.73 \rightarrow 0.38$) and Edge-F1 drops ($0.93 \rightarrow 0.79$). Organ routing ($-0.033$) is next: falling back from the image-to-image neighbour vote to the zero-shot text classifier (94.8% vs. 99.4% organ accuracy) propagates errors through all downstream retrieval, which the organ gates. The remaining choices matter less: report-first over the per-question draft lands within 0.01 of the full method ($-0.008$) on the held-out set. Report-first shows a larger benefit on the official hidden set: in Test Phase 1, it improved Metric A from 0.684 to 0.706 over the separately submitted per-question variant (Section 4.2), motivating its use for the final submission.

### 4.2 REG2026 test-phase results

Table 3 summarises RECAP's performance on the official hidden sets from REG2026 Test Phases 1 and 2. RECAP achieves a Metric A score of 0.706 in Phase 1 and 0.591 in Phase 2. The constant-answer policy of Section 2.4 saturates all three Metric B sub-metrics in both phases (visual grounding 1.0), resulting in weighted overall scores ($0.7\,\mathrm{A} + 0.3\,\mathrm{B}$) of 0.795 and 0.713, respectively.

**Table 3.** RECAP results on the official REG2026 test phases 1 and 2.

| Metric A (workflow) | | | Metric B (grounding) | | |
|---|---|---|---|---|---|
| Component | Phase 1 | Phase 2 | Component | Phase 1 | Phase 2 |
| Binary Path Validity | 0.4514 | 0.3286 | Background rejection | 1.0 | 1.0 |
| Edge-F1 | 0.8320 | 0.7762 | Input sensitivity | 1.0 | 1.0 |
| MESS | 0.7957 | 0.6987 | Cross-region consist. | 1.0 | 1.0 |
| Report score | 0.5883 | 0.4166 | **Visual grounding** | **1.0** | **1.0** |
| **Overall score** | **0.7945** | **0.7134** | | | |

## 5 Discussion and limitations

RECAP relies entirely on retrieval and is computationally cheap, encoding only the test slide at inference and using no generative model in the reporting path, yet it ranks competitively on REG2026. It is best viewed as a strong retrieval baseline rather than a general-purpose reasoning framework. The small gap to naive image-to-image 1-NN indicates that simple case retrieval already accounts for most of the achieved performance. For comparison, SlideChat couples its visual backbone to Qwen2.5-7B-Instruct [4]; its 7B language model alone is roughly 20× larger than RECAP's 355M-parameter WSI backbone. RECAP's quality remains bounded by retrieval-pool coverage and embedding quality, the latter improving as pathology foundation models advance.

**Limitations.** RECAP can handle unseen patients when sufficiently similar diagnostic patterns are represented in the retrieval pool, but its predictions remain constrained by the pool's coverage. Findings that are absent or poorly represented in the retrieval pool therefore cannot be expressed reliably, motivating out-of-distribution detection or identifying low-similarity cases for further review. Furthermore, RECAP provides case-level reasoning but no localized visual evidence linking individual findings to specific WSI regions.

**Future work.** Future work should evaluate the dependence of RECAP on the slide-level representation by comparing TITAN with alternative foundation models such as Prov-GigaPath and PRISM2 [15]. External validation on datasets from different institutions, scanners, and disease distributions is also needed to establish cross-dataset generalisation. Further extensions include flagging out-of-distribution cases for review and linking reasoning steps to supporting tile-level histopathological evidence. RECAP could also serve as the retrieval component of a retrieval-augmented generation framework, combining case-based grounding with the flexibility of generative models.

**On Metric B.** A fixed foreground/background template policy saturates all three Metric B sub-metrics. Beyond distinguishing tissue from background, the answer is independent of both the question and the visual content of the ROI. Thus, a method can achieve a perfect visual-grounding score without grounding its answer in the question or the relevant histopathological features.

## 6 Conclusion

RECAP treats REG2026 pathology reporting as retrieval over frozen foundation model embeddings rather than free-text generation. It requires no task-specific training or language model in the reporting path, while producing a coherent report and reasoning graph from retrieved cases. This simple and computationally efficient design provides a strong baseline for structured whole-slide pathology reporting and can improve directly as the underlying pathology foundation models advance, without task-specific retraining.

## Code availability

The source code used in this work is publicly available at:
`https://github.com/lbuess/reg2026-recap`.

## Acknowledgments

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR @FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG).

## References

1. Ankit Pal, M.S.: Openbiollms: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B` (2024)
2. Buess, L., Perez-Toro, P.A., Bayer, S., Arias-Vergara, T., Maier, A.: Ct-retrievr: fine-grained retrieval-based report generation for 3d ct imaging. In: Medical Imaging 2026: Computer-Aided Diagnosis. vol. 13926, pp. 76–83. SPIE (2026)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
4. Chen, Y., Wang, G., Ji, Y., Li, Y., Ye, J., Li, T., , Ming, H., Yu, R., Qiao, Y., He, J.: Slidechat: A large vision-language assistant for whole-slide pathology image understanding. arXiv preprint arXiv:2410.11761 (2024)
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., et al.: A multimodal whole-slide foundation model for pathology. Nature medicine pp. 1–13 (2025)
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 (HEALTH) **3**(1), 1–23 (2021)
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
8. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature medicine **30**(3), 863–874 (2024)
9. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
10. Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al.: Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024)

11. Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., Vert, J.P.: H-optimus-0 (2024), https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0
12. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
13. Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y.K., Kunz, J.D., Retamero, J.A., Oakley, G., Klimstra, D., Kanan, C., et al.: Prism: A multimodal generative foundation model for slide-level histopathology. arXiv preprint arXiv:2405.10254 (2024)
14. Song, E., Li, H., Bakas, S., Bin Atif, H., Gulzar, M.A., Bappy, Ura, A., Liu, J., Tolkach, Y., Zerbe, N., Khalili, N., Choi, J.H., Schüffler, P., Bilal, M., Ahn, S.: Pathologist reasoning-guided report generation challenge (Apr 2026). https://doi.org/10.5281/zenodo.19848983, https://doi.org/10.5281/zenodo.19848983
15. Vorontsov, E., Shaikovski, G., Casson, A., Viret, J., Zimmermann, E., Tenenholtz, N., Wang, Y.K., Bernhard, J.H., Godrich, R.A., Retamero, J.A., et al.: End-to-end multimodal pathology foundation model with clinical dialogue. Nature Medicine pp. 1–11 (2026)
16. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
17. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (2024)