Faithful Per-Task Attention for Chain-of-Thought Whole-Slide Pathology Report Generation

Gagneet Singh[1], Adinath Madhavrao Dukre[2], and Imran Razzak[1,2]

[1] MedOS, Abu Dhabi, UAE
[2] MBZUAI, Abu Dhabi, UAE

Abstract. Pathology reports can be generated from whole-slide images (WSIs) by following a chain of thought (CoT): before writing the report, a model answers a graph of clinical questions in order. Most multiple-instance learning (MIL) models pool the image patches into a single attention map, then reuse that one map for every question. Methods that split attention do so across the classes of one task, never across different tasks. We remove that limit on a 93-question, 7-organ CoT challenge by giving each question its own attention pool: all 93 heads read the same frozen features, but each learns its own gated attention. Over 3 seeds, at full corpus scale and under the deployed recipe, this improves the structure of the reasoning path with no net loss of accuracy. Masking experiments show the evidence behind each question is its own: removing a question's top-attended patches lowers its predicted probability 4–5 times more, and flips its answer about 3 times more often, than removing another question's. We use *faithful* in a strict sense – the model demonstrably relies on this evidence – which is weaker than matching where a pathologist looks. We also define a Specialization Index, report six standard methods that did not help, and give the full development history, including a segmentation fix worth +0.123 Metric A and our official Test Phase 1 and 2 scores.

Keywords: Computational Pathology · Multiple Instance Learning · Attention Interpretability · Chain-of-Thought Reasoning · Report Generation.

1 Introduction

Most automated pathology reporting maps a whole-slide image (WSI) directly to text. The REG^2 challenge [8] asks for something stricter: the model must first work through a *chain of thought* (CoT), a graph of clinical questions such as "What is the organ?" and "Is there any neoplasm present?", and only then write the report. The score therefore rewards a correct *reasoning path*, not just a correct diagnosis.

The standard method for slide-level prediction is attention-based multiple-instance learning (MIL), where a learned attention distribution pools the patch

features into one slide embedding [2]. CLAM [6] splits this pool into C branches, one per *class* of a single task; TransMIL [11] models how patches relate to each other. Neither splits attention across *tasks*, so for 93 questions the usual choice is still one shared pool – convenient, but hard to justify clinically, since whether a neoplasm is present and how pleomorphic the nuclei are need not be visible in the same region. Recent pathology work shows shared attention can lock onto staining artifacts rather than morphology [5, 1], echoing an NLP debate about whether attention explains anything [3, 12]. Neither asks what happens when attention is untied per task.

Contributions. (1) We give each question its own gated-attention pool over the same frozen features (Eq. 2); shared attention and CLAM are special cases. This improves structural CoT correctness at every scale, with no net loss of accuracy (Sect. 3.2). (2) We show the attention is *faithful* in a causal sense, using masking probes and a new cross-task Specialization Index (Sect. 3.3–3.4). (3) We report the engineering audit behind the submission: a segmentation fix worth +0.123 Metric A, an evaluation formula checked against the official scores, our Test Phase 1 and 2 results, and six standard methods that did *not* help (Sect. 3.1 and 3.5).

2 Method

2.1 Task and decision graph

The challenge provides about 12,000 training H&E WSIs at 20× across seven organs, each with a CAP-style report and a CoT annotation; we use the 11,220 cases that are fully CoT-labeled. The CoT is a closed-vocabulary tree of 93 questions, 21–36 per organ, always starting with "What is the organ?". Each question is a closed-set classification with its own head, and the report is then filled in from 699 templates using the predicted attributes. Given the *ground-truth* answers, the traversal engine reaches a mean edge-F1 of 0.965 but matches the exact path only ≈44% of the time. So the graph logic is essentially solved, what remains is a vision problem, and exact-match scoring is harsh even at this ceiling. A few early "gate" questions (*abnormality present?*, *neoplasm present?*, *histologic type*) decide which branch the model can reach: getting one wrong costs 0.52, 0.44 or 0.23 edge-F1, while every other attribute costs less than 0.08.

Before tiling, a tissue segmenter finds the foreground of each slide, using either a learned model or fast Otsu thresholding; the same choice must be used at training and inference (Sect. 3.1). We tile at 20× and 0.5 MPP into 256 px patches and encode them with H-Optimus-0, a frozen histopathology foundation model [10], using TRIDENT [14], giving $N \approx 500$ to 5000 embeddings of dimension 1536 per slide. The encoder stays frozen so the compute budget goes to the classification heads, the only trained parameters.

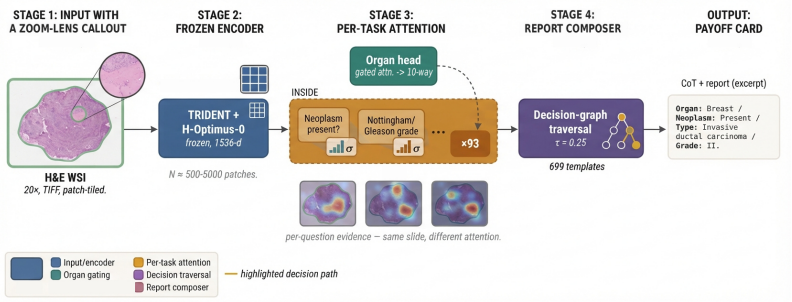

Fig. 1. Pipeline: frozen H-Optimus-0 features feed 93 per-task attention heads (Eq. 2 vs. Eq. 1); predicted answers drive traversal and composition.

2.2 Per-task attention architecture

We illustrate the pipeline in Fig. 1. Gated-attention MIL pooling [2] assigns each patch k, with embedding h_k, a weight

$$a_k = \frac{\exp\{w^\top(\tanh(Vh_k^\top) \odot \mathrm{sigm}(Uh_k^\top))\}}{\sum_j \exp\{w^\top(\tanh(Vh_j^\top) \odot \mathrm{sigm}(Uh_j^\top))\}}, \tag{1}$$

with w, V, U shared across every downstream task, giving one slide embedding $z = \sum_k a_k h_k$ reused by all 93 attribute heads. CLAM [6] generalizes Eq. 1 into C parallel branches w_m, V_m, U_m, one per *class* m of a single slide-level task. We generalize it instead across *tasks*: each attribute head $t \in \{1, \ldots, 93\}$ gets its own parameters w_t, V_t, U_t,

$$a_{k,t} = \frac{\exp\{w_t^\top(\tanh(V_t h_k^\top) \odot \mathrm{sigm}(U_t h_k^\top))\}}{\sum_j \exp\{w_t^\top(\tanh(V_t h_j^\top) \odot \mathrm{sigm}(U_t h_j^\top))\}}, \qquad z_t = \sum_k a_{k,t} h_k, \tag{2}$$

Two known methods fall out of this: setting $w_t = w$, $V_t = V$, $U_t = U$ for all t recovers the shared pool (Eq. 1), and letting t range over the classes of *one* task recovers CLAM. Every question uses the same mechanism, including the 10-way "What is the organ?" that is answered first and gates the rest. We sort the heads into *gate* heads, which decide which branch is reachable; *grader* heads (invasion, Nottingham/Gleason, subtype); *trivial* heads, which are nearly constant; and *compositional* heads, whose outputs are combined by rule into free text.

Implementation. Training minimises a mask-aware sum of per-attribute cross-entropy losses, so each head is trained only on the cases where its question is asked. Class weights are inverse-frequency, multiplied by fixed criticality weights from 1.0 to 6.0, highest for the most important gate. We use AdamW (3×10^{-4}, weight decay 10^{-4}), accumulate gradients over 8 slides per update, train on one A100, and hold out an organ-stratified 18% (seed 0). The deployed recipe caps each organ at 1,600 slides and runs 30 epochs.

2.3 Decision-graph traversal and report composition

At inference the predicted organ selects that organ's question set. Every question the model reaches calls its own head, and a deterministic engine walks

the graph, following a branch whenever the predicted answer's confidence clears an edge-inclusion threshold τ. Sweeping $\tau \in \{0.10, 0.25, 0.35\}$ on the n=140 partition gives 0.829, 0.850 and 0.825, peaking at τ=0.25; Metric A changes by at most 0.025 between the peak and either neighbour, so τ is only mildly sensitive, and an entropy-gated schedule gained nothing (Sect. 3.5). We re-implemented the official formula and use it, not a proxy, for every ablation: Metric A $= 0.05\,\mathrm{BPV} + 0.30$ edge-F1 $+ 0.25\,\mathrm{MESS} + 0.40$ report score, where BPV is binary path validity, MESS is the mean semantic similarity of edge answers, and the report score is $0.40\,\mathrm{J} + 0.30\cos + 0.15\,\mathrm{ROUGE} + 0.15\,\mathrm{BLEU}$ over keyword Jaccard J, Sentence-BERT cosine [9], ROUGE [4] and BLEU [7]. The paper quotes two report scores from different checkpoints. On the 140-case partition the *submitted per-task* model (Table 2, row 3) scores 0.882, with J=0.869, cos =0.946, ROUGE=0.909 and BLEU=0.759; the earlier *shared* checkpoint (row 2) scores 0.82, with 0.811, 0.902, 0.860 and 0.650. Semantic and keyword agreement are strong; exact n-gram BLEU is what lags, as expected for short CAP reports. A typical report matches its reference apart from one numeric field, such as a tumor volume of "50%" instead of "25%". *Interface 0 (Metric B)*, out of scope here, is a classical tissue detector with no learned weights that returns one fixed answer per class. Being deterministic, it makes B_2 and B_3 (0.7 of Metric B) exactly 1 and drives B_1 through a local judge [13], so Metric B saturates at 1.0. Finally, a 140-case audit and a full 11,220-record self-consistency check exposed about 15 composer bugs; fixing them cut a composer-only mismatch from 16.8% to 0.65%.

2.4 Faithfulness protocol and Specialization Index

We use *faithful* in one precise sense, **causal self-consistency**: if we hide the evidence a question attends to, that question's prediction should change more than when we hide another question's evidence or random patches. This tests whether the attention is load-bearing. It does *not* test whether the model looks where a pathologist would look, which would need expert overlays; we measure internal reliance, not human agreement.

We run $n = 800$ probes: 200 held-out slides $\times$ 4 questions. In each we hide the top 15% of patches by attention under three conditions – the question's *own* attention, a *different* question's, and *random* – and record the drop in predicted-class probability and the flip rate. All three remove the same *number* of patches but not the same evidence, so we also record how much own-attention mass each removes and how far the removed sets overlap. The **Specialization Index** measures how different the maps are: per slide, the mean pairwise Jensen–Shannon divergence between different questions' attention distributions, averaged over slides. It is exactly 0 under shared attention by construction (Eq. 1) and an empirical question under per-task attention (Eq. 2). Earlier faithfulness probes [3, 12, 5, 1] work within a single task; this one is defined *across* tasks, and only becomes meaningful once attention is untied per question.

3 Results

All ablations use the official Metric A/B formula on held-out data at pilot (420), full-corpus (11,170–11,220) or 140-slide container scale, with matched seeds $\{42, 7, 123\}$.

3.1 Development history: from a segmentation mismatch to the submitted candidate

Our first Debug-Phase submission trained shared-attention heads (Eq. 1) on features from the learned segmenter, but ran Otsu thresholding at inference to fit the CPU-only runtime limit, breaking the requirement that the two match. It scored Metric A = **0.7055** (Table 2, row 1) with a perfect Metric B. Retraining the same architecture on Otsu features raised Metric A to 0.8286, so the mismatch alone had cost **0.1231**; an audit of the composer and scorer then lifted the report score from 0.7687 to 0.8215, confirmed at 0.8166 on the real 140-WSI container (row 2). The per-task model used a separately extracted feature set, which reintroduced the mismatch, yet row 3 still beats every sub-score in row 2: closing it is upside we do not claim (Sect. 4).

On the official test sets this model scored Metric A = 0.6294 in Phase 1 and 0.5349 in Phase 2, with Metric B = 1.0000 and Totals of **0.7406** and **0.6744** (Table 2, rows 4–5). Both fall below our local validation, and in both the gap is widest on binary path validity (0.33 and 0.23, against 0.61 locally). Two factors unrelated to the model explain most of it. First, the official test set is measurably harder: 19% of its files exceed 1 GB, against 1% in our split, and a fifth arrive pre-pyramidal. Second, binary path validity is an exact-match score over paths averaging 16.6 edges, so BPV $\approx p^{16.6}$ falls away exponentially as per-step accuracy p drops, while partial-credit metrics do not. This is a first-order picture only: it assumes edges fail independently, whereas one early gate failure invalidates the whole path below it, so in practice BPV is set by a few gates. Debug's higher Total is not a regression, since Metric B saturates throughout and Debug was never measured on this harder population, where the same checkpoint reaches 0.897 locally (row 3). The platform returned only aggregate scores, so we cannot report Metric A restricted to normally decoded slides, nor an official fallback count; locally, degenerate composition was rare, at 2 of 1,661 cases (0.12%). Recomputing Metric A from the released Phase-2 sub-scores with our re-implemented formula (Sect. 2.3) reproduces the official Total to four decimal places, independently confirming that our evaluation audit is correct.

3.2 Per-task attention: full-corpus and production confirmation

Table 1 gives each seed's *absolute* score for both conditions (full corpus, 8 epochs, 2,010-slide split) with the mean differences. No seed is worse on any metric, and the means are +0.0017, +0.0067 and +0.0070 for attribute accuracy, edge-F1 and report token-F1. A single production retrain (1,600 slides per organ, 30 epochs, τ=0.25) shows the gain survives the deployed recipe.

We rest the architectural claim on Table 1 alone, where the per-task (PT) and shared (Sh) runs use identical features, composer, split and seeds, so pooling is the only difference. Table 2 is a development history in which segmentation, composer and attention all change from row to row, and the per-task row still carries the mismatch; we use it to explain how the submission was built, never to attribute the gain to the architecture. Its BPV denominator also changes across rows, so the dip in row 2 and the drop in rows 4–5 reflect different populations, not regressions.

Table 1. Per-seed *absolute* held-out scores (full corpus, 8 epochs, 2,010-slide split) for per-task (PT) vs. shared (Sh) attention, with mean 3-seed and production deltas (τ=0.25, 1,660 slides). PT $\geq$ Sh on 8 of 9 cells.

	PT			Sh			Δ	
Metric	s42	s7	s123	s42	s7	s123	3-seed	prod
Attribute acc.	0.883	0.880	0.881	0.879	0.879	0.881	+0.0017	+0.0056
Edge-F1	0.912	0.904	0.911	0.903	0.904	0.900	+0.0067	+0.0130
Report tok-F1	0.889	0.884	0.885	0.878	0.878	0.881	+0.0070	+0.0056

Table 2. Metric A sub-components by development stage (rows 1–3: 140-slide partition; rows 4–5: official Test Phases 1–2, not comparable to rows 1–3). BPV = binary path validity.

Stage	Report	Edge-F1	MESS	BPV
Debug Phase (official)	0.65	0.78	0.73	0.57
+ segmentation & composer fixes	0.82	0.91	0.89	0.55
+ per-task attention (submitted)	**0.88**	**0.94**	**0.93**	**0.61**
Test Phase 1 (official)	0.55	0.75	0.68	0.33
Test Phase 2 (official)	0.42	0.70	0.58	0.23

3.3 Faithfulness is causal, not cosmetic

Hiding a question's own top-attended patches lowers its probability and flips its answer far more than hiding another question's or random ones, tightly across all three seeds (Fig. 2). Per-question attention therefore points at regions the model actually uses for that question, a direct counter-example to the shared-attention failures reported in [5]. Hiding a question's own patches lowers its predicted probability by 0.112, against 0.025 for another question's and 0.003 for random ones, and flips the answer on 13.3% of probes, against 4.5% and 0.9%; the whiskers show the seed range, so the ordering holds for every seed.

Own-masking does remove more evidence: 94.6% of the question's attention mass, against 29.2% and 15.3%. But the removed patches overlap only 24.1% and 15.1% of the time, so the conditions largely delete *different* tissue, and the drop is 4.6× larger while the mass removed is only 3.3× larger (3-seed means, $n = 800$). Figure 3 makes this concrete on one slide. The wrong answer ("invasion present?") has the highest attention entropy (H=0.68) and scattered evidence. The three correct answers each land on one tight region (H=0.22/0.16/0.11).

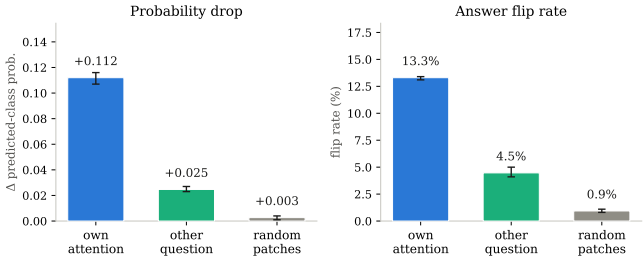

Fig. 2. Causal masking ablation, $n = 800$ probes (200 slides $\times$ 4 questions), 3 seeds (bars = mean, whiskers = seed range); "own"/"other"/"random" = masking the probed question's own vs. a different question's vs. random 15% of patches (Sect. 3.3).

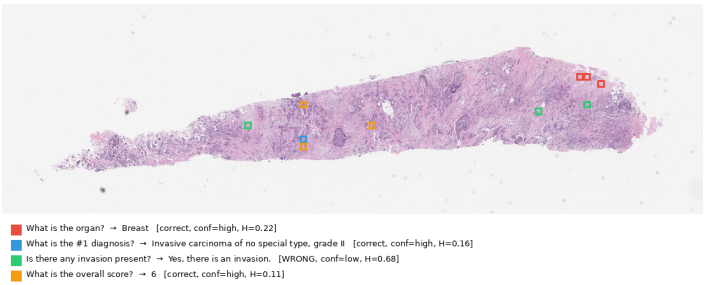

Fig. 3. One illustrative held-out slide: per-question attention evidence (top-3 attended patches, boxed in each question's colour) for four questions; H = normalized attention entropy. The single wrong answer ("invasion present?") carries the highest entropy and the most scattered evidence, while the three correct answers each concentrate on one tight region.

Attention diffusion therefore tracks answer failure question by question, which is impossible under shared attention, where all four inherit one identical map.

3.4 Specialization Index and where the gain concentrates

The Specialization Index (Sect. 2.4) is 0.578, 0.578 and 0.583 for seeds 42, 7 and 123, against exactly 0 for shared attention. That 0 holds by definition, so the bare comparison proves nothing on its own. What matters is that the index's *size* moves with signals measured independently of it: the seeds reaching ≈ 0.58 are the ones with the masking gap and the edge-F1 gain (Table 1), and pushing it *higher* with a diversity loss hurt accuracy (Sect. 3.5). A larger index is therefore not automatically better, but it is not decorative either.

Broken down by head, the gain sits in the gate and descriptive questions ($+0.003$ each) rather than the fine-grained graders (-0.005), which is why edge-F1 moves more than raw accuracy. That is a good trade: graders change edge-F1 by less than 0.08 while the gates change it by 0.52, 0.44 and 0.23, and the -0.005 falls on neighbouring Nottingham and Gleason grades whereas the gains fall on gate decisions that send the path down a different branch. We therefore claim only *no net loss of accuracy overall*, not that everything improved.

3.5 Negative-result screening

Table 3 lists six further standard methods. We screened each under the same seeded, matched-baseline discipline before ruling it out. None transferred.

Table 3. Six standard methods screened under matched controls; none transferred. Metric noted beside each; n = slides.

Method	Effect vs. baseline	Scale (n)
Entropy-gated τ (deployed) edge-F1	−0.0545	1,660
Attention-diversity loss (λ-JS) edge-F1	−0.0017	11,170
Lung label-bug retrain report score	−0.0029	1,661
Confidence-weighted decoding edge-F1	−0.0057	420
Alt. frozen ViT / Macenko stain norm aggregate	≈ 0 (≈0/3 seeds)	420

4 Discussion

Giving each question its own attention pool is a simple change – 93 gated-attention modules instead of one – yet it improves structural CoT correctness at every scale, exactly where the gate sensitivities of Sect. 2.1 predict. The cost is modest: the 93 heads raise trainable parameters from 2.8M to 75M, still only about 7% of the frozen 1.1B-parameter encoder, with nothing added to the backbone and no change in inference latency. That is worth paying because the gain lands on the gates that reroute the whole path rather than being spread thinly, and anyone who disagrees can set all the pools equal and recover shared attention (Eq. 1). Pushing specialization further with a diversity loss did not help, even though it raised the index (Table 3).

Limitations and Future Work. The production retrain is a single run on one A100, differing from the 3-seed mean by +0.0039, +0.0063 and −0.0014. The segmentation fix worth +0.123 Metric A was measured on the *shared-attention* checkpoint, and the submitted per-task model still carries the mismatch. We do *not* assume the two gains simply add: both act on the same features feeding the pooling, so the combined effect may be smaller than their sum, and we leave it unverified. Retraining the per-task heads on matched features is the obvious next step. Our faithfulness claim is bounded in the same way: causal self-consistency shows each head relies on its own evidence, not that this evidence is where a pathologist would look. Testing that needs expert overlays and remains open. Code and weights are publicly available at `https://github.com/adidukre-coder/REG_2026`.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Albuquerque, T., Yüce, A., Herrmann, M., Gomariz, A.: Characterizing the interpretability of attention maps in digital pathology. arXiv preprint arXiv:2407.02484 (2024)

2. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Proc. ICML. pp. 2132–2141 (2018)
3. Jain, S., Wallace, B.: Attention is not explanation. In: Proc. NAACL-HLT. pp. 3543–3556 (2019)
4. Lin, C.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
5. Liu, X., Zhang, W., Tang, W., Le, T., Li, J., Liu, L., Zhang, M.: From correlation to causation: max-pooling-based multi-instance learning leads to more robust whole slide image classification. arXiv preprint arXiv:2408.09449 (2024)
6. Lu, M., Williamson, D., Chen, T., Chen, R., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proc. ACL. pp. 311–318 (2002)
8. REG2026 Challenge Organizers: Pathologist reasoning-guided report generation challenge – task and data description. MICCAI 2026 Satellite Event, OpenReview (2026), `https://openreview.net/group?id=MICCAI.org/2026/Challenge/REG`
9. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proc. EMNLP-IJCNLP. pp. 3982–3992 (2019)
10. Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., Vert, J.: H-optimus-0. Bioptimus (2024), `https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0`
11. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: TransMIL: transformer based correlated multiple instance learning for whole slide image classification. In: Proc. NeurIPS. vol. 34, pp. 2136–2147 (2021)
12. Wiegreffe, S., Pinter, Y.: Attention is not not explanation. In: Proc. EMNLP-IJCNLP. pp. 11–20 (2019)
13. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
14. 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)