This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material Diagnosis-Grounded Structured Decoding for Pathologist Reasoning-Trajectory and Report Generation

Sungyong Jang[1,2], Dongjin Shin[1], Hyunseo Kang[1], and Seung-Won Jung[1]

[1] Korea University, Republic of Korea
[2] AIVIS Inc., Republic of Korea
syjang@aivis.kr

Populations, configurations and estimators

All quantities below are *offline validation* quantities. Test Phase 1 ground truth was never released, so no per-component score on the official test set is computable. Absolute offline levels are optimistic relative to Test Phase 1 (held-out Metric A ≈0.86 against 0.6769 on Test Phase 1); only within-table differences should be read as results.

Populations. VAL-893 is the set of 893 scorable validation slides (the 898 slides of the internal (institution, organ)-stratified slide-level split, minus 5 that lack CONCH features). VAL-789 is the case-disjoint subset of VAL-893: the 789 slides (787 distinct `PIT_<institution>_<case>` prefixes) whose case contributes no slide to any training set used by the pipeline. VAL-789 is the paper's held-out population and is used in Tables S1 and S2; Table S3 is reported on VAL-893 and is labelled as such.

Leak-free artifacts. Every accuracy and score reported here uses diagnosis heads, the retrieval library and the visual k-NN report index rebuilt after removing all 1,022 slides belonging to validation cases (11,220 → 10,198 training slides). The grade heads and the 7-way organ classifier in that artifact set are the deployed (validation-inclusive) weights, and the LoRA generator was trained on the split's training side rather than retrained; these two residual dependencies are shared by every arm, so they affect absolute levels and not within-table differences.

Estimators. Metric A is the official scorer's per-case score

$$0.05\,\mathrm{BPV} + 0.30\,\mathrm{EdgeF1} + 0.25\,\mathrm{MESS} + 0.40\,\mathrm{FinalReport},$$

averaged over the population, with the embedding MESS backend. "Overall" is $0.70\,\mathrm{A} + 0.30\,\mathrm{B}$ with Metric B held fixed at its measured Test Phase 1 value 1.0000: Metric B scores the separate region interface and *cannot* be computed on any offline split, so this is an assumption and not a measurement, and only the

0.70 A part carries sampling variation. The re-weighted columns marked @0.46 apply post-stratification on rank-1 diagnosis correctness,

$$\mathrm{A}@p \;=\; p \cdot \overline{\mathrm{A}}\,[\text{rank-1 dx wrong}] \;+\; (1-p) \cdot \overline{\mathrm{A}}\,[\text{rank-1 dx right}], \qquad p = 0.46,$$

which assumes the two conditional means transfer to deployment and only their mixing weight changes. The value p_{wrong}=0.46 is an assumption with a plausible range of roughly 0.40–0.55, not a measurement. No confidence intervals were computed for the sweeps in Tables S1 and S2: differences below ∼0.003 Overall are within the patient-cluster bootstrap noise of this population and must be read as "flat", not as an optimum.

S1 Sensitivity to the retrieval neighbourhood K

The neighbourhood size K affects only the emitted final report: the reasoning trajectory is the canonical trajectory of the resolved bucket, and it is bit-identical to the deployed setting on all 893 cases for every K we tried. Table S1 sweeps K over $\{1, 3, 5, 10, 20\}$ with everything else held at its deployed value. The whole band is flat and the spread is an order of magnitude smaller than the gap between the free-generation baseline and structured decoding.

Table S1. Sensitivity to the retrieval neighbourhood size K. *Population:* VAL-789 (789 slides, 787 case-disjoint cases). *Configuration:* leak-free diagnosis heads, retrieval library and k-NN index; all other settings at their deployed values (τ_{swap}=0.20, τ_{union}=0.70, consensus tier 1, `REG_HEDGE`=1); uncapped patch bag; official scorer with the embedding MESS backend. *Estimator:* column "Report" is the mean FinalReport term; "A" and "Ovr." are the mean Metric A and $0.70\mathrm{A} + 0.30$ at this population's own held-out rank-1 diagnosis-error rate (p_{wrong}=0.2117, 167/789, in this configuration); "A@.46" and "Ovr.@.46" are the same per-case scores re-weighted to p_{wrong}=0.46 by the post-stratification estimator given in the preamble above. "$\neq K$=5" counts cases whose emitted report text differs from the deployed setting; the non-final trajectory is identical to K=5 on 893/893 cases, i.e. K changes only the final report. **Disclosure:** these sweeps were run with two additional foundation-model views (Virchow2 and Prov-GigaPath) enabled, a four-encoder research configuration that is *not* in the shipped container (`REG_EXTRA_FMS` is empty in `submission/Dockerfile`). Absolute levels therefore sit about 0.009 Metric A above the submitted two-encoder system (0.8648 versus 0.8560 on VAL-789); within-table differences are unaffected. Full-range spread over $K \in \{1, 3, 5, 10, 20\}$, computed from unrounded per-case scores: 0.0023 Overall at the held-out rate and 0.0031 Overall re-weighted to p=0.46. Deployed value starred.

K	Report	A	Ovr.	A@.46	Ovr.@.46	$\neq K$=5
1	0.8166	0.8618	0.9032	0.8096	0.8667	164
3	0.8183	0.8624	0.9037	0.8114	0.8680	104
5*	0.8243	0.8648	0.9054	0.8141	0.8699	0
10	0.8220	0.8639	0.9047	0.8133	0.8693	81
20	0.8248	0.8650	0.9055	0.8136	0.8695	110
Full-range spread: 0.0023 Overall (held-out rate), 0.0031 Overall (p=0.46)						

S2 Sensitivity to the two live gate thresholds

Two continuous thresholds are live in the submitted configuration: the swap threshold τ_{swap}, below which the rank-1 trajectory is replaced wholesale by the rank-2 bucket's, and the union threshold τ_{union}, below which the rank-2 trajectory's unseen edges are added to the rank-1 one. Table S2 sweeps each with the other held at its deployed value and reports how often each tier fires.

Table S2. Sensitivity to the two gate thresholds that are live in the submitted configuration. *Population, configuration, estimator, and the four-encoder disclosure:* identical to Table S1 (VAL-789; leak-free heads/library/index; all non-swept settings at their deployed values; official scorer, embedding MESS backend; held-out rate p_{wrong}=0.2117; @.46 columns post-stratified to p_{wrong}=0.46; absolute levels about 0.009 Metric A above the shipped two-encoder container, differences unaffected). "union" and "swap" are the number of the 789 cases on which the graded edge-union tier and the full rank-2 trajectory swap fire, counted in this four-encoder sweep configuration; for reference, the shipped two-encoder configuration fires swap 137 / union 251 on the same 789 slides and 150 / 266 on all 893. Panel A sweeps τ_{union} with τ_{swap}=0.20 held; the swap tier is evaluated first, so its count is invariant at 114. Panel B sweeps τ_{swap} with τ_{union}=0.70 held; the row τ_{swap}=0.40 lies outside the requested band and is shown only as context. Spreads: Panel A 0.0065 Overall at the held-out rate and 0.0024 at p=0.46 over $[0.5, 0.9]$; Panel B 0.0039 Overall at p=0.46 over $[0.10, 0.30]$. Deployed values starred.

Threshold	A	Ovr.	A@.46	Ovr.@.46	union	swap
Panel A: τ_{union} (graded edge union), τ_{swap}=0.20						
0.5	0.8673	0.9071	0.8126	0.8688	158	114
0.6	0.8665	0.9066	0.8139	0.8698	208	114
0.7*	0.8648	0.9054	0.8141	0.8699	275	114
0.8	0.8618	0.9033	0.8126	0.8688	342	114
0.9	0.8580	0.9006	0.8107	0.8675	440	114
Panel B: τ_{swap} (full rank-2 trajectory swap), τ_{union}=0.70						
0.10	0.8660	0.9062	0.8085	0.8660	318	71
0.15	0.8653	0.9057	0.8105	0.8674	298	91
0.20*	0.8648	0.9054	0.8141	0.8699	275	114
0.25	0.8607	0.9025	0.8119	0.8683	251	138
0.30	0.8568	0.8997	0.8091	0.8664	231	158
0.40 (context)	0.8477	0.8934	0.8052	0.8637	179	210

A third threshold is inert: a code fact, not a robustness result. The variable `REG_CONF_MARGIN_TAU` (deployed value 0.5) has *exactly* zero effect in the submitted configuration, and we state this as a property of the code rather than as evidence of robustness. It is read at `src/infer/structured_decode.py:658` but consumed at two places only: line 678, which sits in the `else:` branch of `if _hedge:`, and the report-backoff helper `_conf_gate_report` (line 451), which

is called only at lines 717–719 under the guard `... and not _hedge`. The submitted container sets `ENV REG_HEDGE=1` (`submission/Dockerfile:165`), so neither consumer is reachable and the parameter is dead code. Sweeping it over $\{0.3, 0.4, 0.5, 0.6, 0.7\}$ reproduces this: every metric is identical to four decimal places (Report 0.8243, A 0.8648, Overall 0.9054, A@.46 0.8141, Overall@.46 0.8699) and the emitted trajectory hash and report text are byte-identical across all five values on 893/893 cases. The same guard makes the two-tier report backoff inert in the submitted container: it fires 0 times on all 893 validation slides and `gate_calls`=0 on all 350 Test Phase 1 slides. We therefore report the two thresholds that are actually live. The one genuinely sharp knob is the discrete consensus tier (`REG_CONF_GATE_THR`): tiers 0 and 1 agree to 0.0011 Overall, while loosening to 2 or 3 costs 0.0378 and 0.0648 Overall@.46; it is fixed at 1.

S3 Per-organ breakdown

Table S3 separates the two decisions the system makes per slide — which organ, and which diagnosis within that organ — and puts each organ's accuracy next to its label cardinality and training support.

Table S3. Per-organ diagnosis and organ recognition. *Population:* VAL-893, i.e. all 893 scorable validation slides — *not* the case-disjoint VAL-789 subset used in Tables S1–S2. On VAL-789 the pooled "dx | true organ" value is 0.7997 rather than 0.8074; per-organ cells were not recomputed on that subset. *Configuration:* validation-case-excluded (leak-free) diagnosis and organ heads, deployed 512-patch bag, shipped two-encoder ensemble, exact raw-label string match. For scale, the same "dx | true organ" measurement with the submitted (validation-inclusive) heads is 0.9317, a 12.4-point memorisation gap. *Estimator:* accuracies are unweighted per-slide proportions; the interval on "dx | organ" is a Wilson 95% interval. "Organ" is the accuracy of the deployed organ path (language-model organ with a visual-classifier override at confidence ≥ 0.95, blocked for cervix); "dx | organ" is rank-1 diagnosis accuracy given the true organ; "Joint" requires both the deployed organ and the rank-1 diagnosis to be correct. Class counts and training support (slides per diagnosis class, as min / median / max, and the number of classes with at most five slides) are counted over the 11,219 labelled slides of the full 11,220-slide training corpus, i.e. before validation-case exclusion; the accuracies use the val-excluded heads.

Organ	n	#dx cls	train slides	min/med/max slides/class	#cls ≤ 5	Organ acc	dx | organ	95% CI	Joint
Bladder	87	9	1079	1/92/294	1	1.0000	0.7241	[0.622, 0.807]	0.7241
Breast	172	52	2212	1/6/730	24	0.9826	0.7500	[0.680, 0.809]	0.7384
Cervix	65	20	812	1/3/357	11	0.9231	0.7077	[0.588, 0.804]	0.6769
Colon	160	24	1996	1/6/575	11	0.9812	0.8125	[0.745, 0.865]	0.7937
Lung	76	16	945	1/18/388	4	0.9737	0.8553	[0.759, 0.917]	0.8289
Prostate	193	4	2418	1/411/1595	2	1.0000	0.9378	[0.894, 0.964]	0.9378
Stomach	140	25	1757	1/7/482	10	0.9571	0.7643	[0.688, 0.827]	0.7286
All	893	150	11219	–	63	0.9787	0.8074	[0.780, 0.832]	0.7917

Reading Table S3. The bottleneck is diagnosis, and it tracks label cardinality and training support rather than organ recognition: prostate has 4 classes with a median of 411 training slides and reaches 0.9378, while breast has 52 classes of which 24 have at most five training slides and reaches 0.7500. Breast, stomach and colon carry 61.6% of the 172 diagnosis errors made given the true organ. Pooled over organs on the same population, diagnosis accuracy against the training support of the ground-truth class is 0.2857 (2/7 cases with support 0–5), 0.4167 (10/24, support 6–20), 0.6387 (76/119, support 21–100) and 0.8520 (633/743, support > 100). The cascade cost of having to predict the organ is small: 0.8074 $\rightarrow$ 0.7917, i.e. -0.0157 or 14 of 893 cases. Standalone organ accuracy on VAL-893 is 0.9709 (867/893) for the language model alone and 0.9877 (882/893) for the visual MIL classifier alone; the deployed combination is 0.9787 because the override is deliberately conservative — it fires on 7 cases, repairs 7 of the language model's 26 organ errors and breaks none.

S4 Full hyperparameter specification

Table S4 specifies the deployed system end to end: patch extraction and encoders, the attention-MIL heads and their training, the multi-view ensembling rule with its per-organ carve-outs, retrieval, the exact gate rules with their thresholds, and the generator with its LoRA and quantisation settings.

Table S4: **Complete hyperparameter specification of the deployed system.** *Population/estimator:* none — this is a static extraction from the shipped checkpoints, the released source and `submission/Dockerfile`, not a measurement. Class counts and the five-seed bagging were read directly from the shipped checkpoints. Values marked ‡ are the committed training-script defaults: the original training invocations were not logged and the checkpoints store only weights and class lists, so those values cannot be verified from the artifacts. The shipped inference encoder is CONCH while the training script's encoder default is H-Optimus, so the real generator run certainly used at least one non-default flag.

Patch extraction and encoders	
Patches	512 px at level 0; kept if tissue fraction ≥ 0.25 under an Otsu saturation mask on a thumbnail of longest side 4096; cap 512 patches/slide (even stride)
Encoders (shipped)	CONCH v1.5 (d=768); H-Optimus-0 (d=1536) on a 224 px *center crop* of the already-read 512 px patch (offset 144; no second read pass)
Encoders (disabled)	Virchow2 (1280), Prov-GigaPath (1536), MedSigLIP (1152) exist in the code but `REG_EXTRA_FMS` is empty in the submitted container

continued on the next page

Table S4, continued.

Attention-MIL heads (identical class for diagnosis, grade and organ)	
Architecture	gated attention: proj Linear(d_{in}→384); V, U Linear(384→256); w Linear(256→1); cls = [LayerNorm(384), Linear(384→K)]
Forward	$h = \tanh(\text{proj}(x))$; $a = \text{softmax}_N\big(w(\tanh(Vh) \odot \sigma(Uh))\big)$; logits = cls$\big(\sum_N a \odot h\big)$; features cast to fp32
Diagnosis heads	150 classes total: breast 52, stomach 25, colon 24, cervix 20, lung 16, bladder 9, prostate 4
Optimiser	AdamW, lr 10^{-3}, weight decay 10^{-4}; no LR schedule, no gradient clipping
Schedule	8 epochs‡; batch = one slide (full bag), one optimiser step per slide, reshuffled each epoch
Loss / bagging	plain cross-entropy, *no* class weighting and *no* resampling; 5 seeds, inference = mean of per-seed softmax
Grade heads	same class and optimiser, 5 seeds, 12 epochs‡; organ skipped if < 50 slides or < 2 classes; classes: prostate 14, breast 12, colon 5, stomach 5, cervix 3, lung 3, bladder 2
Organ head	7-way, d_{in}=768 (CONCH), 5 seeds, 6 epochs‡, lr 10^{-3}, wd 10^{-4}
Multi-view ensembling and its carve-outs	
Rule	unweighted mean of the per-view class posteriors
Diagnosis carve-out	prostate is excluded from the ensemble (CONCH view only)
Grade carve-out	prostate *and* colon excluded; where active the grade ensemble is a fixed 0.5/0.5 mix of CONCH and H-Optimus
Retrieval	
Slide vector	GeM pooling with p=3 over the CONCH patch features, then L_2 normalisation
Similarity / K	cosine; K=5 (`REG_KNN_K`)
Medoid criterion	$\arg\max_i \sum_j \text{Jaccard}(\text{kw}_i, \text{kw}_j)$ over the top-K reports, where kw are scispaCy `en_core_sci_lg` entity keywords; the per-bucket library medoid uses the same criterion over a random sample of at most 40 bucket items (`random.seed(0)`)
Bucket back-off	L3 = (organ, dx, grade signature), requires ≥ 2 items; L2 = (organ, dx), ≥ 1; L1 = (organ), ≥ 1; resolved L3 → L2 → L1
Canonical trajectory	the modal full step signature (question, next_question, answer) within the resolved bucket

continued on the next page

Table S4, continued.

Deployed gate rules (exact)	
Signals	from the ensemble posterior: rank-1 dx, rank-2 dx, margin $= p_1 - p_2$ (probabilities); per-view arg max; consensus = number of per-view diagnoses plus the language model's #1 diagnosis that string-match rank-1 after normalisation (range 0–3; 0–2 on prostate)
Rule 1 (swap)	if consensus ≤ 1 **or** margin < 0.20: full swap to the rank-2 bucket — only if that bucket exists, has a trajectory, and its report passes a 15–4000 character guard; k-NN is re-run inside the new bucket; any exception falls back to rank-1
Rule 1 guard	suppressed if only one view is available (prostate) *and* the CONCH posterior maximum is ≥ 0.9; the union tier is still evaluated
Rule 2 (union)	else if margin < 0.70 and rank-2 $\neq$ rank-1: emit rank-1's steps plus every rank-2 step whose canonicalised (question, next_question) edge key is not already present; the final-report step is forced last and rank-2's final-report step is never added
Rule 3	else emit rank-1 unchanged
Organ override	fires iff the visual classifier's arg max differs from the current organ *and* its probability is ≥ 0.95 *and* the candidate organ has a diagnosis head *and* the candidate is not cervix
Deployed values	`REG_CONF_GATE=1`, `REG_CONF_GATE_THR=1`, `REG_CONF_TRAJ=both`, `REG_HEDGE=1`, `REG_HEDGE_SWAP_TAU=0.2`, `REG_HEDGE_UNION_TAU=0.7`, `REG_HEDGE_GUARD_TAU=0.9`, `REG_KNN_REPORT=1`, `REG_KNN_K=5`, `REG_KNN_SWAP=1`, `REG_ORGAN_CLS=1`, `REG_ORGAN_TAU=0.95`, `REG_NO_OVERRIDE_TO=cervix`, `REG_MAX_PATCHES=512`, `REG_EXTRA_FMS` empty. `REG_CONF_MARGIN_TAU=0.5` and `REG_CONF_BACKOFF=two_tier` are set but unreachable under `REG_HEDGE=1` (Section 3)
Generator	
Backbone / adapter	Qwen2.5-7B-Instruct; LoRA r=32, α=64, dropout 0.05, targets $\{q, k, v, o, \text{gate}, \text{up}, \text{down}\}$_proj, task CAUSAL_LM, `modules_to_save` = {`embed_tokens`, `lm_head`}
Visual front-end	Perceiver resampler, 128 latents, width 1024, depth 4, d_{in}=768 (CONCH)
Training‡	AdamW, peak lr 2×10^{-4}, cosine decay, warmup fraction 0.03, 2 epochs, per-device batch 2, grad-accum 8, grad-norm clip 1.0, max 2048 patches, bf16, report loss weight 1.0 (no report upweighting), `balance_alpha` 0.0 (no class-balanced sampling)

continued on the next page

Table S4, continued.

Quantisation	LoRA merged into the base model, re-saved 4-bit NF4 with double quantisation and bfloat16 compute dtype
Inference	768 new tokens; encode chunk 32 patches; per-case budget 240 s (SIGALRM → catchable `TimeoutError`, degrading to the fallback before the platform's 300 s kill); striped-BigTIFF reader memory cap 12 GB; `PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True`

S5 Interface-0 proxy benchmark

Metric B is computed by the organisers on held-out region-of-interest items that we never see, so it cannot be reproduced offline. Table S5 is an internal proxy for the one decision our Interface-0 implementation actually makes — tissue versus background — comparing the learned classifier we ship against the Otsu/HSV fallback that runs when the classifier checkpoint is absent.

Table S5. Interface-0 tissue/background proxy benchmark. *Population:* 300 ROIs — 150 tissue and 150 background — sampled from official *training* slides only (1024 px level-0 regions resized to 512×512, drawn from 276 opened slides, fixed seed). No Test Phase 1 or organiser-held ROI is involved. *Labelling criterion:* an ROI is labelled tissue if mean HSV saturation ≥ 40 *and* greyscale standard deviation ≥ 20, background if mean saturation ≤ 8 *and* greyscale s.d. ≤ 6; everything in between is discarded. The kept ROIs are therefore unambiguous by construction (tissue: median saturation 57.0, range $[40.3, 95.1]$; background: median 0.6, range $[0.0, 7.3]$), the accuracies below are an easy-case ceiling, and the criterion is not independent of the appearance cues the Otsu arms use — a plain "mean saturation ≥ 15" rule scores 100% on this set. *Configuration:* eight conditions (clean plus seven perturbations) applied to every ROI; three arms — the learned EfficientFormerV2 tissue classifier at probability ≥ 0.5; the **shipped** Otsu/HSV gate, i.e. a verbatim copy of the deployed `src/interf0/model.py::_tissue_fraction` (Otsu on saturation, then zeroing the mask where grey < 20 or grey > 235, then a 3×3 morphological opening) at its deployed threshold 0.03; and a bare-Otsu re-implementation (Otsu on saturation only) at the same threshold. *Estimator:* T is accuracy on the 150 tissue ROIs, B on the 150 background ROIs, and ALL their unweighted mean (balanced accuracy); the aggregate row is the unweighted mean over the eight conditions, which is the figure quoted in the paper (95.9 learned, 95.8 shipped Otsu/HSV, 77.5 bare Otsu). All values are percentages. *Caveat:* this is an internal proxy for the behaviour of the region interface and is **not** a reproduction of the official Metric B components B1 (background rejection), B2 (invariance of the answer under a paired perturbation) or B3 (tissue-versus-background answer difference); those are LLM-judged on organiser-held ROIs and each scored 1.0 on Test Phase 1.

	Learned classifier			Shipped Otsu/HSV gate			Bare Otsu (re-impl.)		
Condition	T	B	ALL	T	B	ALL	T	B	ALL
Clean	100.0	99.3	99.7	100.0	96.0	98.0	100.0	38.7	69.3
Gaussian blur, light (9^2)	97.3	99.3	98.3	100.0	94.7	97.3	100.0	41.3	70.7
Gaussian blur, heavy (31^2)	42.7	100.0	71.3	100.0	94.0	97.0	100.0	47.3	73.7
JPEG, quality 20	100.0	99.3	99.7	99.3	95.3	97.3	100.0	95.3	97.7
Brightness +40	100.0	99.3	99.7	100.0	100.0	100.0	100.0	100.0	100.0
Contrast ×0.5	100.0	99.3	99.7	100.0	64.0	82.0	100.0	41.3	70.7
Saturation ×0.5	100.0	99.3	99.7	100.0	95.3	97.7	100.0	40.7	70.3
Downsample–upsample (128)	98.7	99.3	99.0	100.0	94.0	97.0	100.0	34.7	67.3
Aggregate (mean of 8)	92.3	99.4	**95.9**	99.9	91.7	**95.8**	100.0	54.9	**77.5**

Reading Table S5. The learned classifier and the shipped Otsu/HSV fallback are equivalent on this proxy (95.9 versus 95.8, a 0.1 point difference), so a rebuild without the non-public classifier checkpoint costs essentially nothing here; an earlier internal comparison that found the public fallback materially weaker was in fact measuring the bare-Otsu re-implementation (77.5), not the gate that ships. The two arms fail in different places: the learned model's only weak condition is heavy defocus (71.3, tissue recall 42.7), where it under-calls tissue and the shipped gate is robust (97.0); the shipped gate's weak condition is halved contrast (82.0), where it over-calls background ROIs as tissue. Neither threshold is tuned: pooling all 2400 ROI–condition pairs (a different estimator from the per-condition balanced mean above), the shipped gate scores 95.8% at its

deployed threshold 0.03 and 97.2% at the best swept threshold 0.10, while the bare-Otsu re-implementation peaks at 84.6% (threshold 0.35) and reaches only 77.5% at 0.03.

Provenance note. The 300-ROI benchmark was executed once, in a single run; the code difference between the shipped gate and the bare-Otsu variant, and the mechanism behind it, were verified independently on synthetic near-uniform glass (bare Otsu returns tissue fractions of 0.398–0.430 and calls them tissue, where the shipped implementation returns 0.000). The learned classifier is a pre-existing internal checkpoint rather than a publicly available model; its own training hyperparameters are not available to us.

S6 A diagnosis-compatibility rule for merging candidate trajectories

Referenced from the Discussion. We did not implement or evaluate this rule; it is stated here precisely enough to be implemented and tested.

Partition the closed question pool into *commitment* questions, whose answer fixes a clinical hypothesis (organ; procedure; "Is there any abnormality present?"; "Is there any invasion present?"; "What is the primary of neoplasm?"), and *elaboration* questions, which refine one (histologic type and group, grading system and its component scores, additional findings, the final report). Let T_1 and T_2 be the canonical trajectories of the rank-1 and rank-2 diagnoses. A merge is admissible only if T_1 and T_2 give the same answer, under the scorer's canonicalisation, to *every* commitment question that both contain. When admissible, let d be the earliest question at which their answers differ—necessarily an elaboration question—and add to T_1 only those edges of T_2 reachable from d. Otherwise the merge is refused and the system emits T_1 unchanged.

The emitted graph is then a differential within a single committed behaviour and primary site, a form pathologists do write, while merges spanning benign versus malignant, or two primary sites, are refused outright. The first experiment this calls for is how many of the unions that fired in our submission survive the rule, and what that costs in MESS.

S7 Extending Interface-0 beyond a binary tissue question

Referenced from the Discussion, and offered in response to the reviewer question on the three components of visual grounding. None of this was evaluated in this submission.

Background rejection. Retain the tissue/glass decision as a prior stage, but place behind it a question-conditioned vision–language head that scores answer candidates over several ROIs sampled per slide. Train it with explicit negative-ROI supervision (glass, out-of-focus, pen mark, tissue fold) and an abstention class, so that a non-informative region returns "cannot be determined from this region" rather than a guess.

Input sensitivity. Distinguish morphology-preserving perturbations (stain jitter, compression, brightness and contrast, mild blur, rescaling), under which the answer must not change, from morphology-destroying ones (severe blur, occlusion or removal of the informative region, patch shuffling), under which it must move to abstention. Train a paired invariance/sensitivity objective on these two families rather than reporting a single accuracy, and evaluate them separately.

Cross-region consistency. Aggregate ROI-level answers with a slide-level set or graph model, and apply a consistency term only to questions whose answer is defined at slide level (organ, specimen type, presence of malignancy), not to region-level morphology. Intratumoural heterogeneity is clinically real, so genuine inter-region disagreement should be surfaced as a differential rather than averaged away.

S8 Branch-firing counts and why Test Phase 1 cannot be re-scored per branch

On the held-out population, instrumented replay of the submitted configuration gives: full trajectory swap 137/789, graded edge union 251/789, report backoff 0/789 (on the larger 893-slide validation set, 150/266/0). The report backoff is structurally unreachable whenever the graded hedge is active, which is the submitted configuration.

For Test Phase 1 we replayed the submitted configuration offline over the 350 slides with the shipped artifacts. The report backoff fires on 0/350 exactly. The full swap is bounded at 64–86 slides: the per-slide language-model diagnosis was not persisted at submission time, so 22 slides cannot be resolved without re-running generation, and the edge union accounts for an estimated 153 of the remainder. Official per-branch rescoring is not possible: Test Phase 1 released no ground truth and permitted a single frozen submission, so no counterfactual official score exists for any branch. Every branch contrast in the paper is therefore reported on the held-out population.