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 Decoupling Structure from Perception: A Deterministic-DAG Report Generator with Leak-Fair Cross-Site Recognition for Computational Pathology

Hoang-Vinh Nguyen[0009−0002−8737−025X]

University Medical Center HCMC, Ho Chi Minh City, Vietnam
vinh.nh@umc.edu.vn

1 Organ head: exact training recipe

The organ head is the only supervised component on the structure path (main paper §2.4).

Features. For a bag $X \in \mathbb{R}^{N\times 512}$ of L_2-normalized CONCH-v1 ROI embeddings, the *focal* block is $[\,\mu(X) \,\|\, \max(X) \,\|\, \text{top5}(X)\,] \in \mathbb{R}^{1536}$, where μ is the L_2-renormalized mean, max the coordinate-wise maximum, and top5 the mean of the five highest-norm ROIs. The *anchor* block appends 14 zero-shot margins: for each of the seven organs, an organ-hallmark score (max over that organ's three text prototypes, pooled OR-max over the top-5 ROIs) and a histologic-type score (max over two prototypes, max-pooled over the bag). Only the anchor block is standardized (`StandardScaler` fitted on the training split); the focal block is used raw.

Classifier. Multinomial logistic regression, $C = 0.2$, `class_weight="balanced"`, `max_iter`=5000. Training data are augmented with $K = 5$ Gaussian feature-noise replicas at $\sigma = 0.40$ of each feature's standard deviation (seed 0), which acts as a cheap stand-in for stain augmentation in feature space. At inference the per-class mean training logit is subtracted (per-class debias) – the site-relative operating point of main paper §2.6, computable at deployment without labels.

Software. PyTorch 2.8, timm 0.9.16, scikit-learn 1.5.1, CONCH-v1 weights from the public model card (frozen). Recognition and routing run on a single 16 GB GPU.

Supervision and splits. The label is the organ recorded in the training chain-of-thought; no other annotation is used. Validation is leave-one-site-out: a head is refitted with the evaluation institution entirely excluded, then applied to it. The pooled evaluation basis is six organs on institution B and prostate on institution C ($n = 1{,}657$).

Text prototypes. The prototypes are encoded once with the frozen CONCH-v1 text tower. Organ-hallmark bank (three per organ): bladder – *papillary urothelial fronds lined by umbrella cells*; *multilayered urothelium with surface umbrella cells and no glands*; *urothelial mucosa capped by umbrella cells.* Breast – *specialized loose intralobular stroma of a breast lobule*; *breast terminal duct lobular unit in fibrofatty stroma*; *mammary ducts and lobules.* Cervix – *endocervical mucinous glands at the squamocolumnar junction*; *cervical transformation zone with squamous and columnar epithelium*; *glycogenated stratified squamous epithelium of the uterine cervix.* Colon – *colonic crypts of Lieberkuhn lined by goblet cells*; *Paneth cells with eosinophilic apical granules at the crypt base*; *large bowel mucosa with straight tubular crypts.* Lung – *pulmonary alveoli lined by flat pneumocytes*; *lung parenchyma with alveolar air spaces*; *respiratory bronchial epithelium with cilia.* Prostate – *prostatic acini with luminal corpora amylacea*; *back-to-back prostatic glands with amphophilic cytoplasm*; *prostatic glandular acini.* Stomach – *gastric foveolar epithelium with apical mucin caps*; *deep gastric glands with parietal and chief cells*; *gastric glandular mucosa without goblet cells.*

Histologic-type bank (two per organ): bladder – *papillary urothelial carcinoma with fibrovascular cores and umbrella cells*; *high-grade urothelial carcinoma with pleomorphic transitional cells.* Breast – *invasive ductal carcinoma with tubule formation in desmoplastic stroma*; *infiltrating breast carcinoma cells in single file.* Cervix – *squamous cell carcinoma with keratin pearls and intercellular bridges*; *high-grade squamous intraepithelial lesion with koilocytes.* Colon – *colonic adenocarcinoma with cribriform glands and dirty necrosis*; *colorectal tubular adenoma with pencillate goblet-cell nuclei.* Lung – *lung adenocarcinoma with lepidic alveolar growth*; *keratinizing squamous cell carcinoma of the lung.* Prostate – *prostatic adenocarcinoma with small crowded acini and prominent nucleoli*; *prostate carcinoma with amphophilic cytoplasm and corpora amylacea.* Stomach – *gastric adenocarcinoma with signet ring cells*; *intestinal-type gastric adenocarcinoma with foveolar dysplasia.*

2 Mining the question grammar

Input: the training chain-of-thought corpus. Output, per organ: a node table $\texttt{node} \mapsto \{\texttt{question}, \texttt{kind}, \texttt{next_q}\}$ where `next_q` maps an answer to a *set* of successor questions.

1. Group every CoT step **by case** (not by first occurrence). For a case, a node's answer a is associated with the *set* of questions that follow it in that case.
2. For each (node, answer), store the **modal per-case successor set**. Grouping per case is what preserves ground-truth fan-outs: `grading_system=Nottingham` $\rightarrow$ {tubule formation, nuclear pleomorphism, mitotic count}; `has_abnormality=Yes` $\rightarrow$ {microcalcification, proliferative lesion}; `histologic_type=DCIS` $\rightarrow$ {architecture, nuclear grade, necrosis}. First-occurrence mining collapses each of these to a single edge and loses the diamond.

3. Type each node: *gate* (yes/no question) or *value* (string answer).
4. Key each diagnosis profile by its final diagnosis and store a flat {`node` ↦ `answer`} map plus the report tail and workup bullets.

ROUTE(o, f) then breadth-first walks that table from organ o's root question: read the answer for the current node from f, emit (`question`, `answer`, `next_question`) for *every* successor in the stored set, enqueue those successors, stop at leaves. Because both the questions and the edges are copied from the corpus, no out-of-grammar path can be emitted.

How much grammar is organ-specific? Measured over the training corpus (`_cr_small_numbers.py`): unique questions per organ are breast 36, stomach 31, prostate 28, colon 24, bladder 23, lung 23, cervix 21; only **8** questions are shared by all seven organs (organ, procedure, abnormality, neoplasm-primary, histologic type, number of diagnoses, #1 diagnosis, final report) and **71%** of the corpus's questions occur in a single organ. This is why we did not attempt a single mixed-organ generator as a baseline: it would have to learn per-organ structure that the router copies for free.

3 A worked example: one case end to end

A held-out case (institution B, colon) that the deployed system reproduces exactly. The recognizer turns the 457-ROI bag into an attribute dictionary; the router turns that dictionary into the CoT and the report. Nothing between the two is learned.

```
f = Rec(colon, X):
  organ       Colon                   behaviour       Benign
  procedure   Colonoscopic polypect.  primary      Colorectal
  abnormality Yes                     serration       Yes
  neoplasm    Yes                     histtype        Hyperplastic polyp
  (unreached nodes stay unset: dysplasia, grade, subtype)

c = Route(colon, f):        question | answer | next question
  organ?                  | Colon           | abnormality?
  procedure?              | Colonoscopic p. | abnormality?
  abnormality?            | Yes             | neoplasm?
  neoplasm?               | Yes             | behavior?   <- fan-out
  neoplasm?               | Yes             | primary?    <- same ans
  behavior?               | Benign          | serration?
  primary?                | Colorectal      | serration?
  serration?              | Yes             | histologic type?
  histologic type?        | Hyperpl. polyp  | number of dx?
  number of dx?           | 1               | #1 diagnosis?
  #1 diagnosis?           | Hyperpl. polyp  | final report?
  final report?           | "Colon, colonoscopic polypectomy;
                             Hyperplastic polyp"
```

(Questions are abbreviated to their key phrase; the emitted text is the full corpus wording.)

Two things are visible here. The *fan-out*: one answer (*neoplasm = Yes*) opens two successor questions, both emitted, because the mined edge for that answer is a *set* (§2) – first-occurrence mining would emit only one and lose an edge. And the *failure surface*: every answer in the middle column comes from a recognition probe, so a single wrong answer sends the walk down a different branch; the questions and edges themselves cannot go out of grammar.

4 Site shift: full measurement

Protocol (`_cr_site_shift.py`, log `_cr_site_shift.log`): within an organ, one 512-d bag-mean vector per case from the same frozen coverage embeddings the recognition nodes consume; logistic regression site-vs-site; 5-fold stratified cross-validated AUROC; cosine distance between the two L_2-normalized site centroids.

Table 1. Institution identity is linearly decodable from the frozen embedding in every organ. Control = the same pipeline with the site labels shuffled.

Organ	Pair	n_1	n_2	AUROC / cosine distance
Bladder	A–B	459	146	1.000 / 0.102
Breast	A–B	1248	116	1.000 / 0.070
Cervix	A–B	326	162	1.000 / 0.156
Colon	A–B	418	474	1.000 / 0.190
Colon	A–C	418	328	1.000 / 0.151
Colon	B–C	474	328	1.000 / 0.111
Lung	A–B	272	127	1.000 / 0.234
Prostate	A–C	486	483	1.000 / 0.072
Stomach	A–B	454	149	1.000 / 0.149
Label-shuffled control (7 organs)				median 0.488 $[0.439, 0.582]$

5 Organ misrouting: recovery analysis

Protocol (`_cr_misroute_downstream.py`, log `_cr_misroute.log`): the 62 misrouted cases of the pooled leave-one-site-out evaluation are replayed through ROUTE twice with *ground-truth* attributes – once with the true organ, once with the predicted organ – so the organ grammar is the only difference. Correct-route replay reproduces the ground-truth edge set on 62/62; misrouted replay on 20/62 (32.3%). Correctly-routed control: 1,595/1,595. This is an upper bound on recovery: a real misroute also runs the wrong organ's recognition probes.

Confusion counts among the 62 (true → predicted): stomach→colon 21, lung→bladder 20, lung→stomach 5, lung→cervix 5, colon→stomach 3,

prostate→bladder 2, prostate→lung 2, bladder→breast 1, breast→bladder 1, cervix→colon 1, lung→breast 1. The two dominant confusions are exactly the pairs whose upper grammar coincides, which is why a third of misroutes still produce the correct path.

6 Blind-test score as returned by the organizers

The deployed container was evaluated by the organizers on the Test-Phase-1 hidden institution (submission of 20 July 2026, after the phase leaderboard had closed). Returned metrics: Overall 0.7646; Metric A terms BPV 0.4686, Edge F1 0.8500, MESS 0.7725, Report 0.5341; Metric B (visual grounding) 0.9500, itself background rejection 1.0000, input sensitivity 0.8333, cross-region consistency 1.0000. These reproduce the official weights exactly: $A = 0.05(0.4686)+0.30(0.8500)+0.25(0.7725)+0.40(0.5341) = 0.6852$ and $\text{Overall} = 0.70(0.6852)+0.30(0.9500) = 0.7646$. Input sensitivity is the only Metric-B term below ceiling, and it is the only one a recognition change can move.

7 Cancer detection: the clinical error the challenge metric does not price

Metric A scores paths and text; it does not ask whether a cancer was missed. We therefore compare the malignancy decision of the emitted chain against the ground-truth chain on the same leak-fair held-out folds (`_cr_cancer_confusion.py`).

Table 2. Cancer detection per organ, leak-fair held-out, shipped configuration; Wilson 95% intervals. The intervals carry the caveats: bladder and lung specificity rests on 20 and 7 non-malignant cases, and cervix sensitivity on 11 malignant ones, so those three point estimates say little. Prostate sensitivity is the one estimate that is both low and tight.

Organ	n	TP	FP	FN	Sensitivity [95% CI]	Specificity [95% CI]
Breast	116	72	2	7	0.911 [.828, .956]	0.946 [.823, .985]
Colon	474	67	5	5	0.931 [.848, .970]	0.988 [.971, .995]
Stomach	149	73	12	0	1.000 [.950, 1.00]	0.842 [.744, .907]
Prostate	483	72	13	**30**	**0.706** [.611, .786]	0.966 [.943, .980]
Bladder	146	126	14	0	1.000 [.970, 1.00]	0.300 [.145, .519]
Lung	127	120	4	0	1.000 [.969, 1.00]	0.429 [.158, .750]
Cervix	162	10	1	1	0.909 [.623, .984]	0.993 [.963, .999]
Pooled	**1,657**	540	51	**43**	**0.926** [.902, .945]	0.953 [.938, .964]

The misses are not random. All 30 prostate false negatives are acinar adenocarcinoma, and the colon misses include a well-differentiated adenocarcinoma

called a high-grade adenoma: the system fails on *low-grade, well-differentiated* malignancy – morphologically the hardest kind, and exactly the attribute for which a frozen reader washes out while a trained adapter gains +0.0137 (§3.4 of the paper). Prostate sensitivity 0.706 is the single most clinically consequential number we measured, and it is invisible to Metric A: those cases still contribute edges and report text, so a path-based score barely registers them. Three organs miss no cancer at all, and stomach's zero-FN operating point is bought with 12 false positives – the precision-first setting the paper describes for safety-critical nodes.

8 Node-level failure attribution

For every held-out case whose emitted CoT fails to reproduce the ground-truth edge set, we locate the *first ground-truth edge the prediction does not emit* – the edge, not the answer, since BPV compares edge sets (`_cr_node_attribution.py`). A node whose answer is wrong but whose successor is unchanged leaves the path intact and is counted separately.

Table 3. Nodes that break the path, pooled over the leak-fair evaluation set (n = 1,657).

First unemitted ground-truth edge	% of failures	cases
Is there any abnormality present?	18.6	81
Is there any serration present?	15.2	66
What is the behavior of neoplasm?	14.7	64
Is there any villous component present?	9.7	42
What is the Gleason score?	5.7	25
Is there any invasion present?	5.5	24

Per organ the leading node is breast *#1 diagnosis* (34% of its failures), colon *serration* (36%), lung *adenocarcinoma pattern* (56%), bladder *invasion* (50%), stomach *abnormality* (42%); cervix breaks on 3 of 162 cases. Path-wrong rates run from 1.9% (cervix) to 60.3% (breast).

The analysis also settles a claim the main text otherwise makes by argument: ***procedure* never appears among the path-breaking nodes**. It is wrong on 235 cases, but its successor question does not depend on its answer, so the edge set survives and only the report prefix suffers – exactly the behaviour asserted in limitation (iv).

9 Where the FinalReport points go

FinalReport carries 0.40 of Metric A and is a *text* metric, so a correct diagnosis phrased differently loses points like a wrong one. Each report reads

`<organ>, <procedure>; <diagnoses>`, so splitting at the first semicolon separates the clinical-metadata prefix (organ plus the *defaulted* procedure) from the diagnosis body that the recognizer decides. Over the leak-fair evaluation set (`_cr_fr_decompose.py`):

Table 4. Decomposition of the FinalReport shortfall, $n = 1{,}657$, pooled FinalReport 0.7776.

Case group	n	% cases	mean FinalReport	% of all lost points
Prefix and diagnosis both correct	849	51.2	0.9910	2.1
Prefix wrong only	114	6.9	0.6519	10.8
Diagnosis wrong only	565	34.1	0.5826	**64.0**
Both wrong	129	7.8	0.3387	23.2

Splitting the diagnosis errors further, only 26 cases have a *correct leading diagnosis* with a different enumeration around it, and only 6 are the coexisting-lesion case where we report a second diagnosis and the reference records one (all DCIS); 668 have the leading diagnosis itself wrong. The "penalised for being more sensitive than the annotator" effect is therefore real but small – it is not what holds the score down.

Two readings. (i) **Report construction is not the bottleneck**: when the diagnosis is right the assembled text scores 0.9910, so the template assembler is essentially transparent – the same conclusion the oracle gives for the DAG, arrived at through the text metric instead. (ii) 87.2% of the lost points sit on cases with a wrong diagnosis. The remaining 10.8% is metadata: of 243 prefix errors, 229 are the procedure and only 8 the organ. Making the prefix always correct would lift pooled FinalReport $0.7776 \rightarrow 0.8016$, i.e. Metric A $+0.0096$ – the only headroom in this metric that does not require better recognition.

10 Encoding basis: calibration (600 ROIs) vs deployment (1500 ROIs)

Protocol (`_cr_colon_basis_ab.py`, log `_cr_colon_basis_ab.log`; stomach arm from `_rf1_stomach_*.log`): re-encode a held-out organ on the deployed ROI budget, then re-score the *identical* cases with the identical shipped configuration and the official scorer, so the ROI bag is the only difference.

The effect tracks how far the bag grows, and the two extremes make the mechanism explicit. **Prostate does not move at all**, because a needle-core bag holds 19–103 ROIs: it never reaches either budget, so both configurations return *byte-identical* bags (80/80 cases) and the deployed encoder is simply the calibration encoder. At the other end, bladder and lung roughly 2.3× their ROIs and lose the most. This is what a ratio-thresholded node set does when the denominator grows – concentrated positive evidence dilutes, while gates that

Table 5. All seven organs re-encoded on the deployed ROI budget and re-scored on the identical cases. Five lose points, one is unchanged and one gains, so the calibration-basis numbers of Table 1 read *optimistic* by 0.011 on average. The colon subsample is unbiased on the calibration basis: those 80 cases score 0.7527 where the full 474-case fold scores 0.7528.

Organ (site)	n	median ROIs	calibration	deployment	Δ
Colon (B)	80	259 → 457 (1.76×)	0.7527	0.7593	+0.0066
Prostate (C)	80	68 → 68 (1.00×)	0.9062	0.9062	±0.0000
Cervix (B)	80	235 → 293 (1.25×)	0.9046	0.8944	−0.0102
Breast (B)	80	146 → 250 (1.71×)	0.7474	0.7370	−0.0105
Stomach (B)	149	–	0.8893	0.8765	−0.0128
Lung (B)	80	76 → 176 (2.32×)	0.8581	0.8347	−0.0234
Bladder (B)	80	340 → 782 (2.30×)	0.8873	0.8598	−0.0275
mean					**−0.0111**

were marginally under-filled improve, which is the one organ (colon) that gains. The practical reading: Table 1 is optimistic by ∼0.011 on average (range −0.028 to +0.007) with respect to the container, and a node set thresholded on *ratios* is bounded-but-not-immune to the bag size it is fed. The blind-test score of §6 is unaffected, having been produced on the deployed basis.

A caution for anyone reproducing this. The per-organ bag loaders glob `embs_unified_<organ>_<split>_cov*.npz`. Writing the deployed encodes into the same directory silently makes that pattern match them, so the calibration arm reads the deployment bag and every Δ collapses to exactly 0.0000. The deployed encodes must live in a separate directory; we report the isolated measurement.

11 Procedure node: why it is defaulted

The procedure (biopsy / resection type) is clinical request metadata rather than a morphologic attribute. Measured answer distributions in the training corpus: prostate, bladder and lung have a *single* procedure each (100%); stomach 98.8% (*Endoscopic biopsy*, 4 distinct); breast 98.1% (*Core needle biopsy*, 5 distinct); colon 77.0% (*Colonoscopic biopsy*, 5 distinct, the minority being polypectomy 17.4% and mucosal resection 4.3%); cervix 62.4% (*Colposcopic biopsy*, 6 distinct, punch biopsy 36.5%). The node is therefore defaulted to the modal value per organ; it affects only the FinalReport prefix, not BPV or EdgeF1.

12 Ablation table

Each shipped lever toggled in the deployed configuration, re-scored with the official scorer on its held-out institution. “Fires” counts held-out cases whose emitted CoT changes.

Table 6. Leak-fair ablations (official scorer, ON vs. OFF). Four further shipped levers fire on *zero* held-out cases and are cross-site unproven (§13).

Lever (organ)	fires	Δ A	status
Coexist in-situ 2nd dx, τ=0.99 (breast)	15/116	+0.0176	accepted
same lever, shipped τ=0.80	43/116	−0.0123	shipped (clinical)
Mixed-adenocarcinoma sub-band (stomach)	3/149	+0.0013	accepted
Gleason-grade LoRA adapter (prostate)	–	+0.0137	not shipped

13 Zero-fire levers: what the held-out fold can and cannot show

Four shipped levers change zero held-out cases. A Wilson 95% interval on $0/n$ bounds the *prevalence* of the gated condition, not the lever's effect: breast $0/116 \Rightarrow\leq 3.21\%$; lung $0/127 \Rightarrow\leq 2.94\%$; colon $0/474 \Rightarrow\leq 0.80\%$. Because the event never occurs in the fold, no confidence interval on ΔMetric-A can be formed – which is the point of labelling these levers *cross-site unproven* rather than reporting a calibration-site delta as transfer evidence.

14 Measurement provenance

Table 7. Every camera-ready number added in revision, with its producing script and log.

Claim	Script / log
Site AUROC 1.000, control 0.488, centroid distances	`_cr_site_shift.py` → `_cr_site_shift.log`
Misroute recovery 20/62; control 1,595/1,595; confusion counts	`_cr_misroute_downstream.py` → `_cr_misroute.log`
Wilson bounds; procedure modal shares; grammar overlap 8 / 71%	`_cr_small_numbers.py` → `_cr_small_numbers.log`
Node-level attribution (Table 3)	`_cr_node_attribution.py` → `_cr_node_attribution.log`
FinalReport decomposition; dx-count audit	`_cr_fr_decompose.py`; `_cr_dx_count_audit.py`
Cancer-detection confusion (Table 2)	`_cr_cancer_confusion.py` → `_cr_cancer_confusion.log`
Table 1 per-organ Metric A; ablation deltas and fire counts	official scorer run, `_paper_table1_rerun.log`, `_abl/*.log`
Oracle BPV per organ × site	`_paper_oracle_by_site.py`
Encoding-basis delta: stomach −0.0128; colon +0.0066	`_rf1_stomach_*.log`; `_cr_colon_basis_ab.py` → `_cr_colon_basis_ab.log`

Training and evaluation code, the mined attribute library and a `REPRODUCE.md` were provided to the organizers with the challenge submission.