PATHWEAVER: LEARNING DIAGNOSTIC EVIDENCE GRAPHS FOR PATHOLOGY REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pathology vision-language models (VLMs) increasingly generate diagnostic rationales, yet these rationales are typically expressed as free-form text, making it difficult to verify or supervise individual claims and to trace the chain of evidence that leads to a conclusion. We introduce **PathWeaver**, a framework that makes the diagnostic reasoning behind these rationales structured, explicit, and learnable. It represents each case as a directed acyclic graph (DAG) that links morphological observations, intermediate inferences, and diagnostic conclusions through typed relations. A DAG generator is trained on graphs curated from heterogeneous pathology QA sources and refined by node-level preference optimization, in which counterfactual pairs differ in a single node so that the preference signal localizes to one claim. A graph-conditioned response model then encodes the DAG with relation-aware graph attention to guide answer generation. PathWeaver outperforms general-purpose, medical, and pathology-specialized VLMs across pathology QA formats (66.6% multiple-choice, 55.5% spatial-locating accuracy) and on PathMMU (71.0%). Their explicit structure also enables composition: training-free graph weaving integrates DAGs across patches, improving PanCancer histologic diagnosis benchmark by 10.8% over the best single-view baseline. These results show that an explicit reasoning graph can serve as a single unit of supervision, conditioning, and composition, supporting transparent and inspectable AI-assisted diagnostic reasoning and laying a foundation for more trustworthy collaboration between pathology VLMs and pathologists.

1 INTRODUCTION

Pathology vision–language models (VLMs) are extending computational pathology from fixed prediction tasks to open-ended image–text interaction about histology images by describing morphology, proposing diagnoses, and explaining their answers (Li et al., 2023; Seyfioglu et al., 2024; Dai et al., 2024). Beyond predicting a diagnosis, a growing focus is on enabling pathology VLMs to articulate how they reason from visual findings to diagnostic conclusions. This raises a central challenge: how to structure and explicitly represent the reasoning that connects morphological evidence, domain knowledge, and diagnostic conclusions.

Recent approaches have begun to incorporate reasoning supervision into pathology VLMs. For example, both Patho-R1 (Zhang et al., 2026) and PathReasoner-R1 (Jiang et al., 2026) train on reasoning-oriented data with supervised fine-tuning (SFT), followed by reinforcement learning (RL) to improve diagnostic reasoning. Although natural language can describe these dependencies, individual observations, interpretations, and conclusions remain interwoven within paragraphs, making it difficult to trace. Such dependencies are often many-to-many: a finding may support multiple interpretations, while a conclusion may draw on several findings (Pietarinen & Stanley, 2025; Dalvi et al., 2021). Therefore, explicitly representing these claims and their relationships would enable fine-grained supervision and facilitate their use in downstream answering.

In pathology and broader biomedicine, graph-based representations have been used to organize domain knowledge by encoding entities and their relationships, as exemplified by PathoGraph (Lou et al., 2025) and PrimeKG (Chandak et al., 2023). Structured knowledge also supports pathology model training by providing domain-specific guidance: KEEP incorporates disease knowledge into image–text alignment for disease classification (Zhou et al., 2026), while PathReasoner-R1 uses

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

medical knowledge graphs to guide reasoning-data construction (Jiang et al., 2026). In parallel, intermediate reasoning has been explored through sequential Chain-of-Thought (CoT) rationales (Wei et al., 2022) and Graph-of-Thoughts (GoT), which organizes dependencies among intermediate results to support refinement (Besta et al., 2024). Further discussion of related work is provided in Appendix A. Together, these approaches motivate us to use graphs not only to ground diagnostic claims in structured domain knowledge, but also to guide pathology VLM responses with structured diagnostic evidence.

To this end, we introduce PathWeaver, a framework that learns case-specific diagnostic evidence graphs and uses them to guide pathology question answering (Figure 1). Each graph is a directed acyclic graph (DAG) whose nodes distinguish morphological observations, intermediate interpretations, and diagnostic conclusions, connected through typed relations. A knowledge-grounded curation pipeline transforms heterogeneous pathology data into these graphs, paired textual responses, and localized node-level preference examples. An image–question-conditioned generator first learns to predict the curated graphs and is then refined through node-level preference optimization (Rafailov et al., 2023). The preference objective is restricted to the targeted node, while full-graph teacher-forcing regularization encourages consistency with the accepted diagnostic structure. This makes individual diagnostic claims explicit units of learning without treating them independently of their evidence context.

Our main contributions are threefold:

- **Structured Reasoning.** We develop a knowledge-grounded pipeline that curates diagnostic reasoning as DAGs with typed nodes and edges, with localized counterfactual pairs for claim-level supervision.
- **Learnable and Versatile Graphs.** We train a DAG generator and refine with node-level preference optimization. The learned graphs can extend from single-view answering to multi-patch consensus and exploratory inference without retraining.
- **Explicit Pathology Answering.** We integrate predicted evidence graphs into answer generation through a relation-aware graph encoder, with evaluations demonstrating utility across pathology QA formats.

2 METHODOLOGY

2.1 FRAMEWORK OVERVIEW

Figure 1 provides an overview of PathWeaver, spanning multi-source data curation, explicit diagnostic graph learning and and extended inference capabilities. We first curate diverse pathology QA data into structured reasoning DAGs and corresponding diagnostic responses (Section 2.2). PathWeaver then follows a two-stage architecture comprising a diagnostic DAG generator and a graph-conditioned response model. Given a pathology image x and question q, the generator learns to produce a case-specific DAG that organizes morphological observations, intermediate diagnostic inferences, and conclusions through typed relations (Section 2.3). Node-level preference optimization further learns from medically plausible counterfactuals to localize the preference signal while preserving the global graph format (Section 2.4). The learned DAG is then encoded with relation-aware graph attention and injected into the response model to condition final answer generation (Section 2.5). Finally, a training-free graph weaving strategy integrates DAGs, extending the framework from single-patch to broader-context diagnostic reasoning (Section 2.6).

2.2 DATA CONSTRUCTION: REASONING DAGS AND COT RESPONSES

The training corpus comprises 1,073,803 pathology examples covering multiple QA format including open-ended, dialogue, multiple-choice question (MCQ), spatial localization. Data sources and task-level summary are listed in Appendix B, Table 3. Training our PathWeaver requires reasoning DAGs, corresponding CoT reasoning answers, and node-level counterfactual preference pairs. We construct this supervision through the pipeline shown in Figure 1(a). The LLM-assisted stages use models from the Gemini, Qwen, DeepSeek, and GPT families. Model assignments vary across and processing stages to diversify the generated questions and reasoning responses. Full model versions and assigned tasks are listed in Appendix D.2, Table 7.

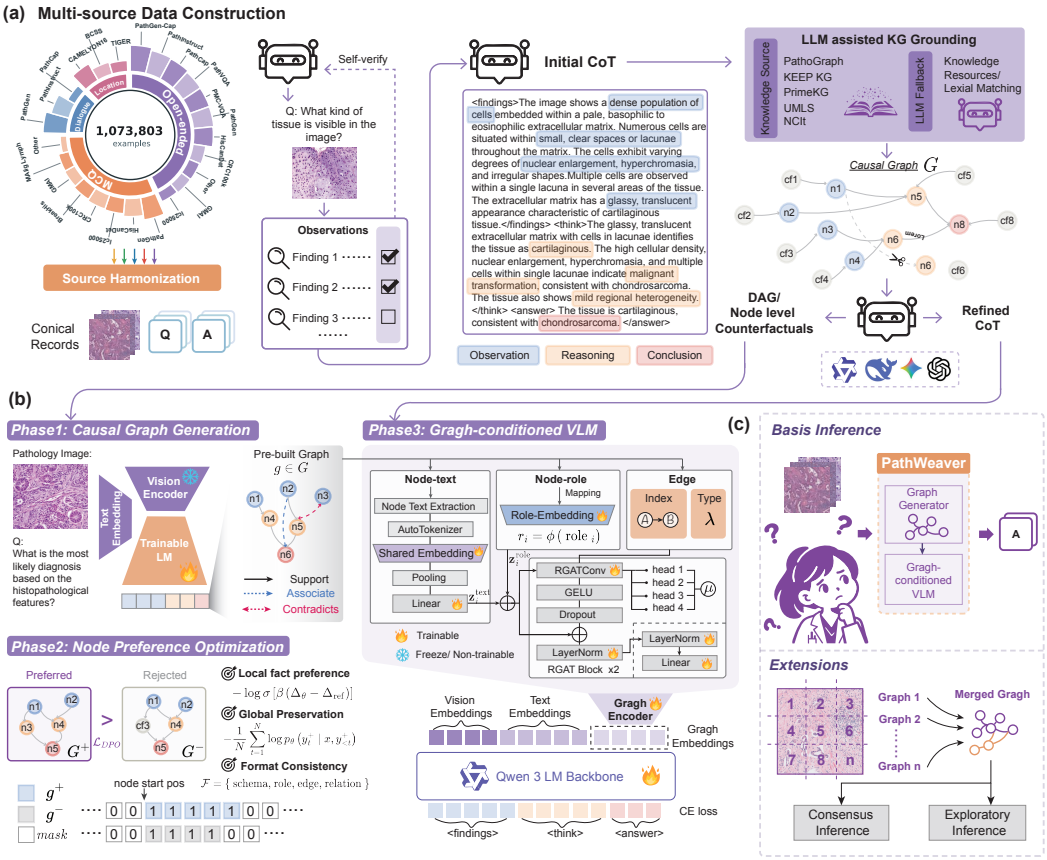

Figure 1: Overview of PathWeaver. (a) Multi-source construction of diagnostic DAGs, reasoning responses, and node counterfactuals. (b) DAG generation, node-level preference optimization, and graph-conditioned response generation. (c) Base inference and training-free extensions.

Source harmonization and initial CoT generation. We begin by converting examples from multiple pathology datasets into a common image–question–answer format, $d = (x, q, a)$. Appendix D details the preprocessing for each source format. For each record, a multimodal LLM generates candidate visual findings $\widetilde{\mathcal{O}} = \{\widetilde{o}_j\}_{j=1}^{J}$ conditioned on the image and question–answer context. Each candidate is then verified against the image, retained or revised when supported by visual evidence, and discarded otherwise, yielding the verified finding set $\mathcal{O}$. The verified findings $\mathcal{O}$, together with the image and question then guide the generation of an initial chain-of-thought (CoT) rationale: $(r^{\text{init}}, a^{\text{init}}) \sim p_{\text{LLM}}(\cdot \mid x, q, \mathcal{O})$. Therefore, each response comprises three components in text: `<findings>`, containing the verified findings $\mathcal{O}$; `<think>`, containing r^{init}; and `<answer>`, containing a^{init}.

Diagnostic DAG construction. An LLM extracts entities from the verified findings $\mathcal{O}$, using the initial rationale and question–answer pair as context. We integrate PathoGraph, KEEP KG, and PrimeKG (Lou et al., 2025; Zhou et al., 2026; Chandak et al., 2023) into a unified knowledge graph $\mathcal{K}$. Entity alignment combines normalized names, type-compatible synonym mappings, and BioBERT semantic matching (Lee et al., 2020), with UMLS and NCIt supporting terminology normalization (Bodenreider, 2004; Sioutos et al., 2007). Evidence-supported paths are then retrieved using entities extracted from $\mathcal{O}$ and $(r^{\text{init}}, a^{\text{init}})$. Since existing knowledge graphs unable to cover all diagnostic reasoning patterns, we use an LLM to complement the retrieved knowledge when beyond the coverage. The LLM then integrates the extracted assertions and retrieved reasoning paths into a diagnostic DAG G. In the final representation, each node is associated with text t_i and a semantic role $\rho_i \in \mathcal{R}_V$, where $\mathcal{R}_V = \{\text{OBSERVATION}, \text{REASONING}, \text{CONCLUSION}\}$ denotes visual observations, intermediate inferences, and diagnostic conclusions, respectively.

Localized preference-pair construction. The diagnostic DAG G comprises a node set V and a typed directed edge set E. Each node v_i pairs text t_i with a semantic role ρ_i:

$$G = (V, E), \qquad v_i = (t_i, \rho_i), \qquad E \subseteq V \times \mathcal{R}_E \times V, \tag{1}$$

where $\mathcal{R}_E$ denotes the set of edge relation types. We use the accepted DAG as $G^+ = G$ and construct a dispreferred counterpart G^- by randomly replacing a single node span with its corresponding LLM-generated counterfactual phrase. Serializing the pair yields $g^+ = \mathrm{ser}(G^+)$, $g^- = \mathrm{ser}(G^-)$ where g^+ and g^- differ only within the perturbed span. Using this procedure, we construct 414,644 training preference pairs for node-level preference optimization (Section 2.4). Appendix E provides construction details.

CoT response refinement. To align the response with the explicit diagnostic reasoning structure, we refine the initial rationale using constructed diagnostic DAG G^+. Formally, the refinement produces revised diagnostic reasoning, and a final answer as:

$$(r^{\mathrm{rev}}, a^{\mathrm{rev}}) \sim p_{\mathrm{LLM}}\big(\cdot \mid q, r^{\mathrm{init}}, a^{\mathrm{init}}, G^+\big)\,. \tag{2}$$

The refinement prompt guides the LLM to improve response coherence and consistency with the structure of G, producing revised reasoning r^{rev} and answer a^{rev} while keeping $\mathcal{O}$ unchanged. Together, these components form the complete supervision response: $y = \mathrm{ser}_{\mathrm{resp}}(\mathcal{O}, r^{\mathrm{rev}}, a^{\mathrm{rev}})$.

2.3 Diagnostic DAG Generator Training

In Phase 1, we supervise a Qwen3-VL-based graph generator p_θ^{DAG} using the image–question–DAG examples constructed in Section 2.2. Given an image x and textual context q, the model generates a JSON-serialized DAG, tokenized using Qwen's byte-level BPE tokenizer (Bai et al., 2025):

$$g = \mathrm{Tok}(\mathrm{ser}(G)) = (g_1, \ldots, g_T). \tag{3}$$

As illustrated in Figure 1(b), we freeze the visual backbone and optimize the visual–language merger and language-model backbone. During training, the target DAG serialization is appended to the input context. We minimize autoregressive cross-entropy over the full DAG serialization, with the input context masked from the loss:

$$\mathcal{L}_{\mathrm{DAG}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta^{\mathrm{DAG}}(g_t \mid g_{<t}, x, q). \tag{4}$$

where θ denotes the trainable parameters, T is the target sequence length, g_t is the target token at position t, and $g_{<t}$ is its ground-truth prefix.

2.4 Node-Level Preference Optimization

Phase 2 optimizes the DAG generator p_θ^{DAG} using the preference pairs (g^+, g^-) constructed in Section 2.2. Both p_θ^{DAG} and the frozen reference model $p_{\theta_{\mathrm{ref}}}^{\mathrm{DAG}}$ are initialized from the same Phase 1 checkpoint. Here, only the language-model parameters are optimized, while the visual backbone and visual–language merger remain frozen. Here, three loss terms are adopted. First, **local fact preference** encourages the model to prefer the original span in g^+ over its counterfactual replacement in g^-. Binary masks $M^\pm$ select tokens overlapping the original and counterfactual spans in $g^\pm$, respectively. For either sequence g and its corresponding mask M, the local score is defined as the mean log-probability over the selected tokens:

$$\ell_\theta(g \mid x, q; M) = \frac{\sum_{t=1}^{|g|} M_t \log p_\theta^{\mathrm{DAG}}(g_t \mid g_{<t}, x, q)}{\sum_{t=1}^{|g|} M_t}. \tag{5}$$

Let $\Delta_\theta = \ell_\theta(g^+ \mid x, q; M^+) - \ell_\theta(g^- \mid x, q; M^-)$, with Δ_{ref} computed using the reference model. The local objective follows the Direct Preference Optimization (DPO) formulation (Rafailov et al., 2023):

$$\mathcal{L}_{\mathrm{local}}(\theta) = -\log \sigma(\beta[\Delta_\theta - \Delta_{\mathrm{ref}}])\,, \tag{6}$$

where σ is the sigmoid function and $\beta = 0.1$ scales the preference margin. Second, **global preservation** retains full-graph supervision through $\mathcal{L}_{\mathrm{global}}(\theta)$, defined as the graph-generation loss in

Equation 4 evaluated on the preferred sequence g^+. Third, **format consistency** applies additional supervision to structural tokens in g^+, selected by a binary mask:

$$F_t = \begin{cases} 0, & \text{if token } g_t^+ \text{ overlaps node text values,} \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

Thus, the mask selects node identifiers, semantic roles, edge endpoints, relation labels, and JSON syntax, while excluding node text values. The format loss averages cross-entropy over tokens with $F_t = 1$:

$$\mathcal{L}_{\text{format}}(\theta) = -\frac{1}{\sum_{t=1}^{T_+} F_t} \sum_{t=1}^{T_+} F_t \log p_\theta^{\text{DAG}}(g_t^+ \mid g_{<t}^+, x, q). \tag{8}$$

With $\gamma_{\text{global}} = \gamma_{\text{format}} = 0.1$, the overall objective is:

$$\mathcal{L}_{\text{NPO}}(\theta) = \mathcal{L}_{\text{local}}(\theta) + \gamma_{\text{global}}\mathcal{L}_{\text{global}}(\theta) + \gamma_{\text{format}}\mathcal{L}_{\text{format}}(\theta), \tag{9}$$

2.5 Graph-Conditioned Response Training

In Phase 3, we initialize a separate graph-conditioned response model p_ψ^{resp} from a pretrained Qwen3-VL checkpoint (Figure 1(b)). Training uses pre-built diagnostic graphs, whereas inference uses those generated by the preference-optimized DAG generator.

The DAG encoder integrates node text, semantic roles, and typed edges through three branches. The **node-text branch** embeds tokenized node texts using the response model's shared input embedding matrix E_ψ, followed by mean pooling over non-padding tokens and a trainable linear projection. The **node-role branch** provides a learned role embedding $\phi(\rho_i)$. Their outputs are added to initialize each node:

$$h_i^{(0)} = \text{Linear}_{\text{text}}(\text{MeanPool}(E_\psi(\text{Tok}(t_i)))) + \phi(\rho_i). \tag{10}$$

The **edge branch** supplies edge endpoints and relation labels from E, with $\mathcal{R}_E = \{\text{SUPPORTS}, \text{ASSOCIATED_WITH}, \text{CONTRADICTS}\}$. Given the integrated node representations $H^{(\ell)} = [h_1^{(\ell)}, \ldots, h_{|V|}^{(\ell)}]$, relational graph attention (RGAT) aggregates incoming neighbor features using relation-specific transformations and attention weights. We apply two RGAT blocks with four heads averaged in each layer (Busbridge et al., 2019). Each block updates the node representations as:

$$H^{(\ell+1)} = \text{LN}_\ell\Big(H^{(\ell)} + \text{Dropout}\Big(\text{GELU}\Big(\text{RGAT}_\ell(H^{(\ell)}, E)\Big)\Big)\Big), \quad \ell \in \{0, 1\}. \tag{11}$$

An output LayerNorm and linear projection map the final representations into the language model's hidden space:

$$z_i = \text{Linear}_{\text{out}}\Big(\text{LN}_{\text{out}}(h_i^{(2)})\Big), \qquad Z_G = [z_1, \ldots, z_{|V|}]. \tag{12}$$

Here, z_i denotes the representation of node i contextualized by relation-aware neighborhood aggregation. Graph delimiters and one placeholder per node are appended to the question, where each placeholder is replaced by the corresponding z_i for conditioning the response model. With this conditioning, the response model is trained on the refined CoT target y constructed in Section 2.2 by minimizing token-averaged cross-entropy:

$$\mathcal{L}_{\text{resp}}(\psi) = -\frac{1}{T_y} \sum_{t=1}^{T_y} \log p_\psi^{\text{resp}}(y_t \mid y_{<t}, x, q, G), \qquad T_y = |y|. \tag{13}$$

At inference (Figure 1(c), top), the generator first produces $\widehat{g} \sim p_\theta^{\text{DAG}}(\cdot \mid x, q)$ and then extract $\widehat{G} = \text{parse}(\widehat{g})$; the response model then generates $\widehat{y} \sim p_\psi^{\text{resp}}(\cdot \mid x, q, \widehat{G})$.

2.6 Cross-Patch Graph Weaving

Cross-patch graph weaving extends single-patch reasoning by integrating diagnostic evidence across tissue regions within a relative large image. For tissue microarray (TMA) cores as example, tissue regions can be sampled on a grid to obtain patches $\{x_p\}_{p=1}^P$ after CLAM's tissue-segmentation pipeline (Lu et al., 2021). As illustrated in Figure 1(c), PathWeaver further supports cross-patch

compositional reasoning, in which the DAGs $\widehat{G}_p$ can be predicted independently for each grid under a shared question q, then woven into a single graph. On their disjoint union $G_{\sqcup} = \bigsqcup_{p=1}^{P} \widehat{G}_p$, node texts are encoded using a frozen SapBERT encoder f_{merge} (Liu et al., 2021):

$$\widetilde{z}_w = \frac{f_{\text{merge}}(t_w)}{\|f_{\text{merge}}(t_w)\|_2}, \qquad s_{uv} = \widetilde{z}_u^{\top} \widetilde{z}_v \tag{14}$$

where $\widetilde{z}_w$ is the L_2-normalized embedding of node w and s_{uv} measures cosine similarity between the text embeddings of nodes u and v. Two nodes are merged if their canonicalized texts match exactly, or if $s_{uv} \geq \tau_{\text{merge}}$ and negation and qualifier checks pass (encoder selection and threshold sensitivity are reported in Appendix K.1). Each resulting group is collapsed into one node that takes the text of its most central member (the one with the highest mean s_{uv} to the others) and the majority role. Remapping the edges accordingly and dropping self-loops and duplicates yields $\overline{G} = (\overline{V}, \overline{E})$, which is finally pruned to the nodes that support a conclusion or contradict such support:

$$v^{\star} = \arg \max_{v \in \overline{V}_{\text{con}}} N_{\text{patch}}(v) \tag{15}$$

where $\overline{V}_{\text{con}}$ is the set of retained conclusion nodes and $N_{\text{patch}}(v)$ counts the patches whose original DAG contains a supports path from an observation to a member of v; ties follow stored node order. Results on TMA data are reported in Section 3.2.

3 Experiments

PathWeaver instantiates both models from Qwen3-VL-4B/8B, giving four configurations denoted generator /response model (e.g., 8B/4B); the training corpus and algorithms are described in Appendices B and F. We compare against general-purpose, medical, and pathology-specialized VLMs (listed in Appendix A); models fine-tuned on our own supervision without graphs share PathWeaver's training data and serve as controls in the whole-core evaluation (Section 3.2) and the ablations (Section 3.4). The evaluation sets and metric definitions and the LLM-as-a-judge protocol are given in Appendix C and G.

3.1 Pathology Question Answering

The held-out internal evaluation covers three QA formats: 500 multiple-choice, 500 open-ended, and 449 location questions (Appendix C.1). Table 1 (left) reports multiple-choice accuracy, along with GPT-5.4-judged open-ended correctness (0–5) and location accuracy (%). BLEU is additionally reported as a reference-based metric. Definitions of all metrics and the evaluation protocols are provided in Appendix G. All four PathWeaver configurations outperform the baselines in Table 1 (left) on every reported metric. The 8B/4B and 8B/8B configurations achieve 66.60% multiple-choice accuracy, while 4B/8B achieves the highest open-ended correctness of 2.978. PathWeaver (8B/8B) achieves 55.46% location accuracy, compared with 28.06% for InternVL3.5-8B, the strongest baseline on this metric. In addition, evaluation on the public pathology QA benchmark PathMMU (Sun et al., 2024a) covers 9,580 questions across five subsets, with subset-level and overall accuracy reported in Table 1 (right). PathWeaver (8B/8B) achieves the highest overall accuracy of 70.95%, compared with 63.62% for Patho-R1. PathWeaver also achieves the highest accuracy on each subset, with the leading configuration varying by subset. Training-stage comparisons are provided in Section 3.4.

3.2 Consensus Inference via Cross-Patch Graph Weaving

A single crop of a tissue microarray (TMA) core captures only part of the tissue, so core-level questions call for consensus inference over patch-level evidence. We evaluate it on the PanCancer benchmark (2,466 questions across ten cancer types; Appendix C.2) using the same weaving settings (Section 2.6) for all cancer types, without organ-specific tuning. At inference, the woven DAG replaces the single-patch graph as the condition of the response model, which changes the answer in two ways: (1) the graph pools observations from every patch, so the response draws on evidence beyond a single field of view, and (2) it retains only the conclusion supported by the most patches (Eq. 15), so the response commits to a consensus rather than to one patch's verdict. The node encoder f_{merge}, which decides whether two node texts refer to the same concept, is selected on a curated benchmark of 1,000 annotated node-text pairs, where SapBERT outperforms seven other biomedical text encoders in accuracy, merge F1, and ROC-AUC, as shown in Figure 2(a,b); the benchmark

Table 1: Performance comparison on the QA format test sets and PathMMU; a/b denotes DAG generator/response model sizes. Baseline references are given in Appendix A.

	MCQ	Open-ended		Location		PathMMU					
Model	**Acc.↑**	**BLEU↑**	**Corr.↑**	**BLEU↑**	**Acc.↑**	**Atlas↑**	**Edu-Content↑**	**Path-CLS↑**	**Pub-Med↑**	**Social-Path↑**	**Overall↑**
General-purpose open-source VLMs											
LLaVA-1.5-7B	35.40	0.041	1.330	0.121	16.48	37.14	34.42	21.17	36.31	36.35	33.16
LLaDA-MedV	35.00	0.023	1.510	0.066	6.01	38.83	32.09	20.12	33.31	32.54	31.01
InternVL3.5-8B	52.20	0.045	2.356	0.074	28.06	58.09	57.33	39.75	57.01	58.25	54.15
InternVL3-14B	54.60	0.057	2.330	0.057	15.14	60.77	58.57	32.78	58.60	59.50	54.11
Qwen3-VL-8B	48.40	0.015	2.162	0.031	22.27	40.32	39.58	40.41	42.89	44.31	41.74
Gemma 3-12B	42.60	0.012	2.286	0.024	14.92	37.04	39.16	34.22	40.09	39.31	38.33
Medical and pathology-specialized baselines											
MedGemma 1.5-4B	43.40	0.030	2.242	0.072	5.57	48.16	48.35	31.84	46.97	47.33	44.58
MedGemma-27B	46.40	0.029	2.278	0.039	5.12	44.59	40.09	26.26	40.32	40.84	38.16
LLaVA-Med-v1.5-7B	33.40	0.069	1.962	0.189	8.02	32.57	28.38	19.62	29.20	27.70	27.31
PA-LLaVA	39.80	0.002	1.236	0.006	1.78	42.50	40.30	28.47	38.46	38.91	37.45
Quilt-LLaVA-v1.5-7B	43.00	0.035	1.692	0.102	18.49	49.75	45.46	26.59	40.32	45.39	40.69
Patho-R1-7B	55.80	0.040	2.026	0.120	14.48	75.67	68.47	43.45	67.11	66.04	63.62
PathWeaver Family											
PathWeaver (4B/4B)	64.20	0.190	2.968	0.372	53.45	76.27	70.85	54.12	74.97	70.25	69.47
PathWeaver (8B/4B)	**66.60**	0.181	2.956	**0.381**	54.34	75.37	74.46	53.51	76.40	**72.64**	70.89
PathWeaver (4B/8B)	65.40	**0.193**	**2.978**	0.375	51.00	**76.56**	71.00	**54.17**	75.07	70.08	69.54
PathWeaver (8B/8B)	**66.60**	0.187	2.954	**0.381**	**55.46**	75.37	**74.61**	53.51	**76.53**	72.58	**70.95**

construction and evaluation protocol are described in Appendix J. The weaving hyperparameters were selected on a separate 1,216-question development set and are listed in Appendix K.

On PanCancer, the gains from graph conditioning and from weaving can be read separately in Figure 2(c). On the same thumbnail input, graph conditioning alone lifts Qwen3-VL-8B from 63.18% (graph-free SFT) to 71.41% (+8.23%), surpassing the best single-view baseline (Patho-R1, 66.75%). Weaving patch-level graphs then adds a further +2.52%, reaching 73.93% for the 8B/8B configuration; the gain from weaving holds for all four DAG generator/response model configurations (Table 11, Appendix H.3). To test whether this gain comes merely from seeing more tissue, we apply majority voting over the same patches to the graph-free 8B SFT model, which reaches only 68.94% (Appendix I). Figure 2(d) shows a representative case in which the woven graph consolidates evidence from several patches into the correct diagnosis of squamous cell carcinoma, missed by all compared baselines. The response cites a longer evidence chain than the baselines, and every cited finding maps to a node of the merged graph, so the answer is verifiable claim by claim.

3.3 Exploratory Inference: Whole-Slide Case Study

A whole-slide image can span thousands of patches, making it challenging to identify and localize relevant findings across the slide. A single consensus conclusion cannot answer such a question, and no VLM can read a gigapixel slide in one pass. Exploratory inference (Section 2.6) merges semantically equivalent nodes from patch-level DAGs into a unified graph. A selected conclusion is then sementically matched to local nodes to produce a concept-localization map, and the graphs from the highest-scoring regions can provide traceable evidence for response generation. The held-out CAMELYON16 (Ehteshami Bejnordi et al., 2017) example in Figure 3 illustrates the query guided localization followed by evidence-supported responses on a sentinel lymph node slide. The map for "metastatic carcinoma" closely aligns with the annotated tumor region (patch-level AUC of 0.93), while the maps for lymphoid and adipose tissue highlight complementary regions consistent with the expected tissue morphology. We further evaluate on the CAMELYON17 dataset (Bandi et al., 2018) (details in Appendix M). The resulting localization maps distinguish tumor from other area, achieving mean per-slide ROC-AUCs of 0.90 for macrometastases and 0.71 for micrometastases. Further cases from this set is shown in Appendix N. Beyond localization accuracy, our maps link highlighted regions to explicit diagnostic rationales, with supporting evidence traceable through the unified graphs. Heatmaps derived from attention weights and gradient-based saliency heatmaps can highlight relevant regions (Lu et al., 2021; Selvaraju et al., 2017), while the diagnostic rationale remains implicit (Lou et al., 2026). The displayed patches are randomly sampled from high-scoring

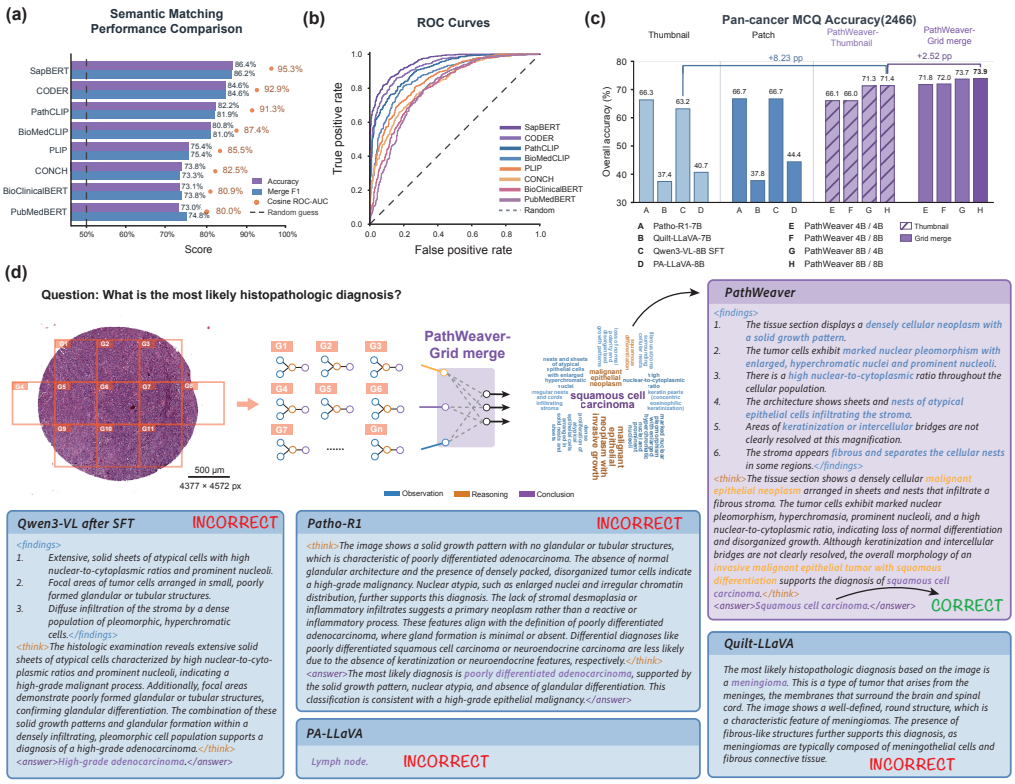

Figure 2: Node canonicalization and whole-core graph weaving. (a–b) Node canonicalization performance, (c) PanCancer MCQ results under single-view and graph-weaving settings, and (d) a qualitative example of whole-core graph-grounded reasoning.

candidates for each resulting conclusion, where patches within each group exhibit similar morphology (Figure 3, middle; Appendix N). The interaction is thus interpretable at the level of individual claims, with each localization traceable through the reasoning graph to its corresponding morphological evidence and reasoning path. This makes localization an auditable process rather than an output that must simply be trusted.

3.4 Ablations and Analysis

Table 2: Training-stage ablation of PathWeaver across pathology QA benchmarks.

Model / config.	PathMMU ↑	MCQ ↑	Location ↑	OpenQ ↑
Graph-free fine-tuning (vs. base models in Appendix L, Table 16)				
LLaDA-MedV SFT	61.54 (+30.53)	57.60 (+22.60)	40.76 (+34.75)	2.776 (+1.266)
Qwen3-VL-4B SFT	67.37 (+30.61)	62.60 (+14.00)	44.10 (+27.17)	2.808 (+0.604)
Qwen3-VL-8B SFT	70.25 (+28.51)	62.80 (+14.40)	48.55 (+26.28)	2.864 (+0.702)
Qwen3-VL-8B SFT+GRPO	69.97 (+28.23)	64.60 (+16.20)	47.66 (+25.39)	2.878 (+0.716)
PathWeaver: DAG SFT-only (vs. Qwen3-VL SFT, matched size)				
PathWeaver (4B/4B)	67.25 (−0.12)	62.60 (+0.00)	46.10 (+2.00)	2.900 (+0.092)
PathWeaver (4B/8B)	67.21 (−3.04)	62.00 (−0.80)	46.10 (−2.45)	2.874 (+0.010)
PathWeaver (8B/4B)	69.93 (+2.56)	64.00 (+1.40)	44.10 (+0.00)	2.952 (+0.144)
PathWeaver (8B/8B)	69.89 (−0.36)	64.40 (+1.60)	43.65 (−4.90)	2.972 (+0.108)
PathWeaver: Full (vs. DAG SFT-only, matched configuration)				
PathWeaver (4B/4B)	69.47 (+2.22)	64.20 (+1.60)	53.45 (+7.35)	2.968 (+0.068)
PathWeaver (4B/8B)	69.54 (+2.33)	65.40 (+3.40)	51.00 (+4.90)	**2.978** (+0.104)
PathWeaver (8B/4B)	70.89 (+0.96)	**66.60** (+2.60)	54.34 (+10.24)	2.956 (+0.004)
PathWeaver (8B/8B)	**70.95** (+1.06)	**66.60** (+2.20)	**55.46** (+11.81)	2.954 (−0.018)

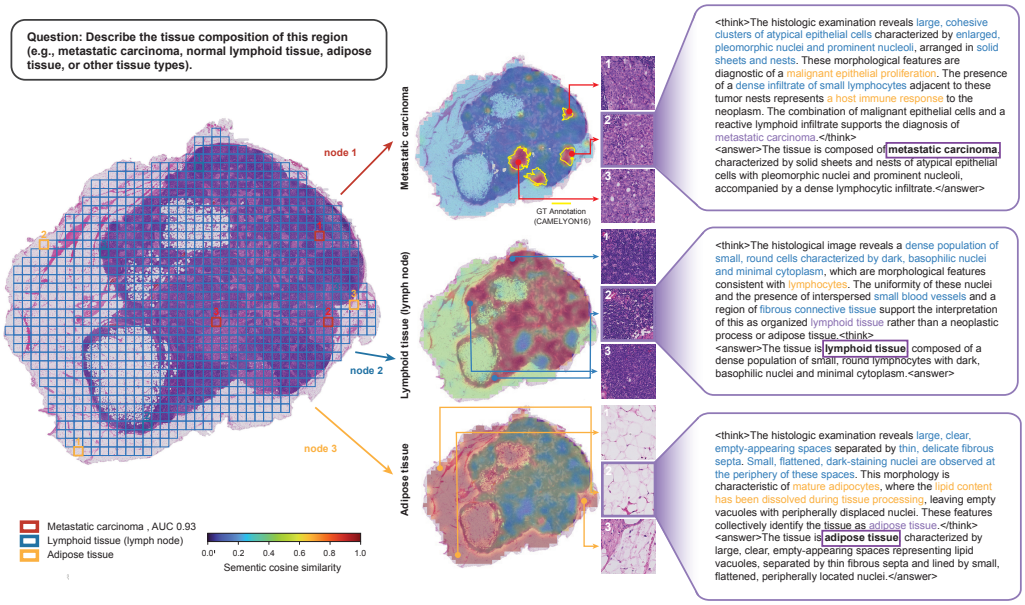

Figure 3: Exploratory whole-slide inference example on CAMELYON16

PathWeaver differs from a fine-tuned VLM in two respects, graph conditioning and node-level preference optimization, and the ablation asks how much each contributes. Table 2 therefore reads top to bottom as the training pipeline, with gray values giving matched-size changes from the previous stage (Appendix L). Graph-free fine-tuning on our supervision lifts every base model by 14–35%. Conditioning on graphs from an SFT-only generator (DAG SFT-only) then changes results only modestly and inconsistently (-3.0 to $+2.6$ on PathMMU, -4.9 to $+2.0$ on location), whereas adding node-level preference optimization (Full) improves every configuration on the closed-form and location benchmarks: by 0.96–2.33% on PathMMU, 1.60–3.40% on multiple choice, and 4.90–11.81% on the location subset, with open-ended correctness within ± 0.1. The location subset benefits most, consistent with the counterfactual pairs targeting factual errors in individual nodes. Relative to the graph-free 8B SFT model, full PathWeaver (8B/8B) gains 3.8% on multiple choice, 6.9% on location, and 0.7% on PathMMU. Across configurations, performance is more sensitive to the DAG generator than to the response model: moving the generator from 4B to 8B raises PathMMU by about 1.4% and PanCancer by about 1.9%, whereas changing the response model at a fixed generator shifts results by at most 0.21% (Table 11), suggesting that graph quality is the primary bottleneck.

4 Conclusion

PathWeaver introduces case-specific diagnostic graphs as a shared representation for localized supervision, response generation, and cross-patch evidence integration. Knowledge-grounded graph construction and node-level preference optimization make individual diagnostic claims explicit targets of learning, while graph-conditioned answering uses these claims and their relationships to guide responses. Experiments demonstrate improvements across pathology QA tasks and highlight the contribution of node-level preference optimization. Training-free graph weaving extends the use of these graphs to whole-core consensus inference and query-guided whole-slide exploration. Together, these findings support explicit reasoning graphs as a way to connect claim-level learning with evidence-guided answering and make intermediate reasoning available for inspection.

The evaluation remains limited by the additional model capacity and inference cost of graph generation, which are not controlled in comparisons with graph-free models. Future work should address this computational trade-off and improve the robustness of graph weaving. Beyond benchmark performance, explicit evidence graphs could support clinical workflows by helping pathologists review diagnostic claims, examine the observations linked to them, and question model-generated reasoning. This points toward more transparent human–AI collaboration, in which AI-assisted evidence integration complements the contextual expertise and diagnostic judgment of pathologists.

AI USE STATEMENT

We used large language models (e.g., GPT-6) for non-substantive assistance during manuscript preparation. Specifically, LLMs were used to improve writing clarity, grammar, and phrasing, but not for generating scientific content or experimental design. All technical contributions, experiments, and interpretations were conceived and conducted by the authors.

The authors take full responsibility for the content of the manuscript, including any text generatedor polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guide lines and does not contribute to plagiarism or scientific misconduct.

ETHICS STATEMENT

This study primarily uses publicly available pathology datasets. The PanCancer benchmark comprises de-identified clinical data collected under institutional ethics approval. Institutional details are withheld for double-blind review. The PanCancer data are currently subject to licensing restrictions and will be made publicly available once these restrictions are lifted. Dataset sources and cohort details are documented in Appendices B and C. PathWeaver's outputs are intended as a supplementary reference and should not replace pathologists' clinical judgment.

REPRODUCIBILITY STATEMENT

All code, training scripts, and instructions necessary to reproduce our results are available at the anonymized repository: https://anonymous.4open.science/r/pathweaver-252A/

REFERENCES

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pp. 72–78, 2019.

Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18): 3461–3467, 2019.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.

Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y Hammerla. Relational graph attention networks. *arXiv preprint arXiv:1904.05811*, 2019.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific data*, 10(1):67, 2023.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37: 94327–94427, 2024.

Dawei Dai, Yuanhui Zhang, Long Xu, Qianlan Yang, Xiaojing Shen, Shuyin Xia, and Guoyin Wang. Pa-llava: A large language-vision assistant for human pathology image understanding. In *2024 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pp. 3138–3143. IEEE, 2024.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7358–7370, 2021.

Xuanzhao Dong, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Peijie Qiu, Shao Tang, Xin Li, and Yalin Wang. Llada-medv: Exploring large language diffusion models for biomedical image understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22773–22783, 2026.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Camelyon16 Consortium, Meyke Hermsen, Quirine F Manson, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

Jiarun Fu, Lizhong Ding, Hao Li, Pengqi Li, Qiuning Wei, and Xu Chen. Unveiling and causalizing cot: A causal pespective. *arXiv preprint arXiv:2502.18239*, 2025.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183. IEEE, 2024.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

Songhan Jiang, Fengchun Liu, Ziyue Wang, Linghan Cai, and Yongbing Zhang. Pathreasoner-r1: Instilling structured reasoning into pathology vision-language model via knowledge-guided policy optimization. *arXiv preprint arXiv:2601.21617*, 2026.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 31038–31065, 2023.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 4228–4238, 2021.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Peiliang Lou, Yuxin Dong, Caixia Ding, Chunbao Wang, Ruifeng Guo, XiaoBo Pang, Chen Wang, and Chen Li. Pathograph: A graph-based method for standardized representation of pathology knowledge. *Scientific Data*, 12(1):872, 2025.

Peiliang Lou, Yitan Zhu, Nicholas Chia, Roopa Kumari, William Yang, Yan Wang, Brenna C Novotny, Stacey J Winham, Ruifeng Guo, Ellen L Goode, et al. A human-in-the-loop explanation framework for morphologically transparent ai predictions from whole-slide images. *npj Digital Medicine*, 2026.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

Nikita V Orlov, Wayne W Chen, David Mark Eckley, Tomasz J Macura, Lior Shamir, Elaine S Jaffe, and Ilya G Goldberg. Automatic classification of lymphoma images with transform-based global features. *IEEE Transactions on Information Technology in Biomedicine*, 14(4):1003–1013, 2010.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Ahti-Veikko Pietarinen and Donald E Stanley. Diagnostic reasoning in surgical pathology. *Journal of evaluation in clinical practice*, 31(5):e70219, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Mart van Rijthoven, Witali Aswolinskiy, Leslie Tessier, Roberto Salgado, Jeroen van der Laak, and Francesco Ciompi. Analysis of computational tumor-infiltrating lymphocytes in breast cancer from the results of the tiger challenge. *Nature Communications*, 2026.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13183–13192. IEEE, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Nicholas Sioutos, Sherri de Coronado, Margaret W Haber, Frank W Hartel, Wen-Ling Shaiu, and Lawrence W Wright. Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, 40(1):30–43, 2007.

Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63 (7):1455–1462, 2015.

Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, et al. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, pp. 56–73. Springer, 2024a.

Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5034–5042, 2024b.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Jingxiong Li, Xuan Gong, Xinheng Lyu, Tao Lin, et al. Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In *International Conference on Learning Representations*, volume 2025, pp. 94611–94653, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218. Springer, 2018.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025). *URL https://arxiv. org/abs/2508.18265*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, 126:103983, 2022.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arxiv 2023. *arXiv preprint arXiv:2303.00915*, 70, 2023a.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Wenchuan Zhang, Penghao Zhang, Jingru Guo, Tao Cheng, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28418–28426, 2026.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023b.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277, 2024.

Xiao Zhou, Luoyi Sun, Dexuan He, Wenbin Guan, Ge Wang, Ruifen Wang, Lifeng Wang, Xiaojun Yuan, Xin Sun, Ya Zhang, et al. Knowledge-enhanced pretraining for vision-language pathology foundation model on cancer diagnosis. *Cancer Cell*, 44(4):777–791, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

A RELATED WORK

Pathology VLMs and diagnostic rationales. Open general-purpose VLMs (LLaVA-1.5, InternVL3/3.5, Gemma 3, Qwen3-VL) (Liu et al., 2024; Zhu et al., 2025; Wang et al.; Team et al., 2025; Bai et al., 2025) are the common starting point and remain competitive on pathology benchmarks even without domain adaptation. Domain adaptation has mostly followed visual instruction tuning. For example, LLaVA-Med (Li et al., 2023) applies it to biomedical image–text data, MedGemma (Sellergren et al., 2025) builds on broad clinical pretraining, and pathology-specific models such as Quilt-LLaVA trained on narratives extracted from educational histopathology videos (Seyfioglu et al., 2024) and PA-LLaVA on pathology QA pairs curated from PathVQA and PMC-VQA (Dai et al., 2024; He et al., 2020; Zhang et al., 2023b). Diffusion language models have recently emerged as an alternative to autoregressive decoding (Nie et al., 2026), and LLaDA-MedV brings this formulation to biomedical images with the same instruction tuning strategy (Dong et al., 2026). More recent work has shifted toward optimizing the reasoning process: Patho-R1 and PathReasoner-R1 supervise models with pathology reasoning traces and further optimize their reasoning behavior through reinforcement learning (Zhang et al., 2026; Jiang et al., 2026; Shao et al., 2024). In all of these models, however, supervision stays at the level of the entire response, and the evidential dependencies among claims remain implicit in the text. PathWeaver instead represents claims and their dependencies as an explicit graph that serves as both a training target and a conditioning input for guiding answer generation.

Reasoning graphs and knowledge graphs. Chain-of-thought prompting elicits a linear sequence of reasoning steps (Wei et al., 2022), and Graph-of-Thoughts generalizes it to a graph whose vertices are partial results that can be combined and refined (Besta et al., 2024); causal-graph prompting similarly extracts a graph of variables or reasoning steps to guide inference (Jin et al., 2023; Fu et al., 2025). In these methods the graph is constructed at prompting time and is not itself a training target. In pathology, structured knowledge exists in resources such as PathoGraph (Lou et al., 2025), PrimeKG (Chandak et al., 2023), the NCI Thesaurus (Sioutos et al., 2007), and the UMLS (Bodenreider, 2004); KEEP uses disease knowledge to align pathology images and text (Zhou et al., 2026), and PathReasoner-R1 uses knowledge graphs to guide the construction of reasoning data (Jiang et al., 2026). These works use a fixed knowledge graph as an external guide.

Localized preference optimization. Preference learning such as DPO (Rafailov et al., 2023) compares two complete responses. Finer variants attach the signal to parts of a linear chain of thought: process reward models score individual reasoning steps (Lightman et al., 2024), Step-DPO contrasts responses at the first erroneous step (Lai et al., 2024), and TDPO formulates the preference objective at the token level (Zeng et al., 2024). PathWeaver applies the same idea to a graph. Each preference pair differs in a single node of the diagnostic DAG and the loss is restricted to that node (Section 2.4), so the model is told which finding or inference is wrong, not merely which response is worse.

B Training Corpus Composition

Table 3: Composition of the harmonized training corpus by task type. Counts are examples after source harmonization.

Training task	Examples	Proportion	Data sources
Open-ended QA	833,101	77.58%	PathGen-caption; PathInstruct; PathCap; PathVQA; PMC-VQA; PathGen-Instruct; His Can Det; GMAI-MMBench; lc25000; CRC100k; BreakHis; MAlig Lymph
Multi-turn dialogue	146,214	13.62%	PathGen-Instruct; PathInstruct; PathCap
Multiple-choice QA	83,271	7.75%	PathGen-Instruct; His Can Det; lc25000; CRC100k; BreakHis; GMAI-MMBench; MAlig Lymph
Region-grounded QA	11,217	1.04%	CAMELYON16; BCSS; TIGER
Total	1,073,803	–	15 source datasets

Table 3 summarizes the corpus obtained after the source harmonization stage of Section 2.2. 15 public pathology sources are mapped into canonical image–question–answer records and grouped into four training tasks. During curation, we exclude non-pathology samples with insufficient visual quality, and images substantially obscured by annotation overlays or other visual artifacts, resulting in 1,073,803 examples.

Open-ended question answering accounts for the largest proportion of the dataset (77.58%), caption-style dataset such as PathGen-caption and PathCap contribute a substantial portion of these examples. Multi-turn dialogue (13.62%) and multiple-choice questions (7.75%) provide additional forms of supervision: dialogue data captures multi-step reasoning over a shared image, whereas some multiple-choice data originated from classification datasets (lc25000, CRC100k, BreakHis). Region-grounded questions (1.04%) are constructed from segmentation and detection annotations in datasets like CAMELYON16, BCSS, and TIGER, providing explicit spatial supervision that is evaluated on the location subset in Section 3.1.

We describe each source below and how it contributes to the QA tasks in our harmonized corpus:

- **PathGen-caption.** The image–caption component of PathGen-1.6M contains pathology patches paired with descriptions generated and refined through multi-agent collaboration. We convert these descriptions into open-ended QA examples (Sun et al., 2025).
- **PathGen-Instruct.** Released alongside PathGen-1.6M, this instruction-following dataset contains pathology questions and image-based conversations. It contributes both multiple-choice QA and multi-turn dialogue (Sun et al., 2025).
- **PathCap.** Released with PathAsst, PathCap contains pathology image–text pairs collected from scientific publications and educational resources. It contributes primarily to open-ended QA, with a small number of records categorized as multi-turn dialogue (Sun et al., 2024b).
- **PathInstruct.** PathInstruct provides pathology instruction–response data constructed as part of PathAsst. We harmonize its records into open-ended QA and multi-turn dialogue (Sun et al., 2024b).
- **PathVQA.** PathVQA pairs pathology images with questions about their visual content. We retain the question–answer structure and normalize the records into our open-ended QA format (He et al., 2020).
- **PMC-VQA.** PMC-VQA contains medical visual questions derived from figures and accompanying text in scientific publications. We use its pathology subset as open-ended QA (Zhang et al., 2024).
- **Histopathologic Cancer Detection (His Can Det).** This dataset contains lymph-node histology patches labeled for metastatic tissue detection and is a modified version of PatchCamelyon. Its associated QA records contribute binary classification content to our multiple-choice task (Veeling et al., 2018; Hu et al., 2024).

- **LC25000.** LC25000 contains lung and colon histopathology images spanning benign tissue and malignant categories. Its associated multiple-choice questions provide tissue and tumor classification supervision (Borkowski et al., 2019; Hu et al., 2024).
- **CRC100k.** NCT-CRC-HE-100K contains colorectal histology patches annotated with nine tissue classes. Its associated multiple-choice questions provide supervision for distinguishing tumor and surrounding tissue components (Kather et al., 2019; Hu et al., 2024).
- **BreakHis.** BreakHis contains benign and malignant breast tumor images acquired at multiple magnifications. Its associated QA records contribute breast histopathology classification questions to the multiple-choice task (Spanhol et al., 2015; Hu et al., 2024).
- **MAlig Lymph.** A malignant lymphoma dataset of H&E images labeled by subtype (chronic lymphocytic leukemia, follicular lymphoma, and mantle cell lymphoma). We convert the subtype labels into multiple-choice and open-ended QA about lymphoma classification (Orlov et al., 2010).
- **GMAI-MMBench.** A multimodal medical benchmark spanning many imaging modalities and clinical tasks. We use its pathology subset, converting the multiple-choice items into multiple-choice QA and their reference answers into open-ended QA (Chen et al., 2024).
- **CAMELYON16.** CAMELYON16 provides lymph-node whole-slide images with reference annotations of breast cancer metastases. We use these annotations to construct region-grounded questions about the presence and location of metastatic tissue (Ehteshami Bejnordi et al., 2017).
- **BCSS.** The Breast Cancer Semantic Segmentation dataset provides breast histology images with tissue-region annotations. We use its category labels and spatial masks to construct region-grounded questions(Amgad et al., 2019).
- **TIGER.** TIGER provides breast cancer histology data with tissue-region and cell annotations for assessing tumor-infiltrating lymphocytes. We use its spatial annotations to construct region-grounded QA(Rijthoven et al., 2026).

C Evaluation Datasets

C.1 Internal single-view test sets.

The three held-out test sets are constructed from source data reserved for evaluation before curation, following the same QA construction procedure described in Appendix D. These data are excluded from all model-training stages. The multiple-choice and open-ended sets each contain 500 questions sampled uniformly at random from the corresponding task, spanning all contributing source datasets. The location set is derived from the region-grounded questions, whose answers originate from annotation masks and bounding boxes (Appendix D) provided from source dataset. From initial 500 randomly sampled questions we retained the 449 whose reference answers state an explicit location, so that the correctness of the stated location can be judged(Appendix G).

C.2 Pan-cancer test dataset

The Pan-cancer test dataset is an in-house benchmark built from H&E-stained tissue-microarray (TMA) cores covering ten cancer types. All cores are held out from training. The data were collected with institutional ethics approval and fully de-identified.

Table 4 lists the histologic diagnosis questions: 2,466 multiple-choice questions over 2,433 TMA core images from 1,550 patients across ten cancer types, which form the PanCancer benchmark evaluated in Section 3.2 and reported per organ in Appendix H.3. The ten cancer types are represented almost uniformly, with 100–300 questions each.

Table 4: Histologic diagnosis questions in the Pan-cancer test dataset by cancer type, with the number of patients and TMA core images they cover.

Cancer type	Patients	Core images	Questions
Prostate	63	115	115
Cervical	126	244	258
Liver	151	254	260
Colorectal	197	260	260
Lung	158	266	266
Kidney	143	242	255
Gastric	187	265	265
Pancreatic	227	268	268
Bladder	69	247	247
Esophageal	229	272	272
Total	**1,550**	**2,433**	**2,466**

An additional development set of 1,216 whole-core questions is constructed using the same protocol as PanCancer, with no patient-level overlap with the PanCancer test set. This set is used to select graph-weaving hyperparameters, as detailed in Appendices K.1 and K.2.

C.3 PathMMU benchmark

PathMMU (Sun et al., 2024a) is a public, expert-validated multiple-choice benchmark for pathology image understanding and reasoning. Its questions are drawn from five sources with different image and question styles: PubMed (figures from the biomedical literature), SocialPath (cases shared by pathologists on social media), EduContent (educational videos and slides), Atlas (textbook atlases), and PathCLS (classification-style questions derived from public patch-level datasets). We adopt it as the external benchmark because it is the largest expert-validated pathology QA test set, covers image sources disjoint from our training corpus, and is the standard evaluation for recent pathology VLMs, including the reasoning-oriented models compared in Table 1. We evaluate on the official test split. Of its 9,677 questions, 97 questions corresponding to 62 SocialPath images were excluded because the original publicly hosted images could no longer be retrieved, leaving 9,580 evaluable questions; all models in this paper are evaluated on the same 9,580 questions, and Table 5 gives the number of evaluable questions per subset.

Table 5: Evaluable questions per subset in the PathMMU test split used in this paper.

Subset	Questions
Atlas	1,007
EduContent	1,938
PathCLS	1,809
PubMed	3,068
SocialPath	1,758
Total	**9,580**

D Source Harmonization

Source harmonization (Section 2.2) converts each raw input format into a common image–question–answer representation. Let r_f denote a raw example with format f. A format-specific mapping H_f produces the canonical record

$$H_f(r_f) = d = (x, q, a),$$

where x, q, and a denote the image, question, and answer, respectively. The image is retained from the source record, whereas the question and answer are copied, normalized, or constructed according to the available annotations. After harmonization, every record enters the same graph-curation pipeline.

Table 6 distinguishes information provided by a source from information constructed during harmonization. Open-ended VQA is already in canonical form. For multiple-choice data, the option set $\mathcal{C}$ is appended verbatim to the question, written as $[q; \mathcal{C}]$, and $\mathrm{Norm}(y; \mathcal{C})$ converts the answer to a fixed option-label format. For a dialogue $\mathcal{T}$, we prepend the preceding turns c_T to the final question q_T, forming the question input $[c_T; q_T]$, and use the final response a_T as the answer.

Caption sources require both fields to be constructed. Given a caption s, the question and answer fields are generated sequentially as $q_{\mathrm{cap}} = \mathrm{LLM}_Q(s)$ and $a_{\mathrm{cap}} = \mathrm{LLM}_A(s, q_{\mathrm{cap}})$, yielding the open-ended record $(x, q_{\mathrm{cap}}, a_{\mathrm{cap}})$.

For spatial annotations, m denotes the annotation mask, c the associated category labels, and $\mathcal{B}$ the region boxes extracted from the mask. The programmatic mapping Ψ constructs a supervision answer from the selected category and boxes.

Table 6: Format-specific mappings from heterogeneous source annotations to the canonical record (x, q, a). LLM_Q and LLM_A denote question and answer generation where shown. A check mark denotes a source field retained without further generation.

QA format	**Raw record**	**Question** q	**Answer** a	**Resulting record**
Open-ended	(x, q, a)	✓	✓	(x, q, a)
Multiple-choice	$(x, q, \mathcal{C}, y)$	$[q; \mathcal{C}]$	$\mathrm{Norm}(y; \mathcal{C})$	$(x, [q; \mathcal{C}], \mathrm{Norm}(y; \mathcal{C}))$
Multi-turn dialogue	$(x, \mathcal{T})$	$[c_T; q_T]$	a_T	$(x, [c_T; q_T], a_T)$
Image caption	(x, s)	$q_{\mathrm{cap}} = \mathrm{LLM}_Q(s)$	$a_{\mathrm{cap}} = \mathrm{LLM}_A(s, q_{\mathrm{cap}})$	$(x, q_{\mathrm{cap}}, a_{\mathrm{cap}})$
Spatial annotation	(x, m, z)	$q_{\mathrm{ann}} = \mathrm{LLM}_Q(x, m, z)$	$a^{\mathrm{sup}} = \Psi(q_{\mathrm{ann}}; z, \mathcal{B})$	$(x, q_{\mathrm{ann}}, a^{\mathrm{sup}})$

D.1 Spatial-Annotation QA Construction

For CAMELYON16, BCSS, and TIGER, we use spatially annotated histology images, with binary tumor masks for CAMELYON16 and multiclass tissue masks (including tumor, stroma, lymphocytic infiltrate and necrosis) for BCSS and TIGER. We apply 8-connected component labeling to category-specific masks, filter small components by area, and merge nearby bounding boxes to obtain a compact set of region boxes for each category. These boxes are retained as spatial annotations for generating QA pairs. Given the image, a colorized annotation mask, and category label, the multimodal LLM generates questions about the presence, identity, or location of one tissue components. So that the labels and region boxes can provide the initial supervision answers.

D.2 Models Used for Data Curation

Table 7 summarizes the specification of used models, their roles, and official references.

Table 7: Models used for data curation, their assigned stages, and official references.

Model	Curation stages	Official reference
Gemini 3.1 Flash-Lite	Question generation; finding generation; finding verification; initial CoT; CoT refinement	Model card
Qwen3.5-397B-A17B-FP8	Question generation; caption-to-QA; finding generation; finding verification; initial CoT; CoT refinement; concept normalization; DAG construction; counterfactual generation	Model card
DeepSeek-V4-Flash	Caption-to-QA; concept normalization; text-only CoT refinement; DAG construction; counterfactual generation	Documentation
GPT-5.4	Question generation; finding generation; finding verification; initial CoT; CoT refinement	Release page

E Localized Preference-Pair Construction

We validate graph structure and candidate spans before constructing preference pairs. During DAG preprocessing, we check acyclicity using topological sorting. A graph passes this check only if all nodes can be assigned positions in an ordering π satisfying:

$$\pi : V \to \{1, \ldots, |V|\}, \qquad \pi(v_i) < \pi(v_j) \quad \text{whenever } v_i \to v_j. \tag{16}$$

Here, $\pi(v_i)$ denotes the position of node v_i in the ordering, so each node precedes its successors.

The example prompt for node-aligned counterfactual phrase generation is shown below.

Counterfactual candidate-generation prompt

```
Create pathology counterfactual replacements for all three
response sections using the reasoning DAG.

Rules:

- Cover every DAG node that has an exact compact phrase in the
  corresponding section.
- Each replacement must align to a different node within its
  section.
- ‘original‘ must be an exact substring of that section.
- ‘counter‘ must be a grammatical drop-in medical or pathology
  replacement that substantially changes the meaning.
- ‘counter‘ must be mutually exclusive with the original claim
  for the same target and context.
- Do not use a synonym or paraphrase as the counter.


Return JSON only:
{"find": [{"dag_node_id":"n1","original":"...","counter":"..."}],
 "think": [...], "answer": [...]}

Question: {question}

Reasoning DAG:
{dag_text}

Findings:
{findings}
```

```
Think:
{think}

Answer: {answer}
```

F TRAINING ALGORITHMS

Algorithms 1 and 2 summarize Phases 2 and 3; Phase 1 follows standard SFT (Section 2.3), with the loss computed only over output tokens; input-context and padding tokens are excluded.

For each DAG serialization $g^{\pm}$, $s^{\pm}$ records the start and end character positions of the edited span (Appendix E). The mask $M^{\pm}$ selects tokens overlapping that span and must be nonzero.

Algorithm 1 Node-level preference optimization of the DAG generator

Require: Preference records $(x, q, g^+, g^-, s^+, s^-)$; SFT checkpoint
Require: Weights β, γ_{global}, γ_{format}
Ensure: Preference-optimized DAG generator p_θ^{DAG}

```
Initialize policy and reference models from the same SFT checkpoint
Freeze the reference model and the policy's visual backbone and merger
for each minibatch of preference records do
    for each record (x, q, g+, g−, s+, s−) do
        Tokenize g+, g− while retaining character offsets
        Construct M+, M− by aligning tokens with s+, s−, respectively
        Construct F on g+ using Equation 7
        Compute policy token log-probabilities for both sequences given (x, q)
        Compute reference token log-probabilities without gradient tracking
        Evaluate L_local, L_global, and L_format using Equations 6, 4, and 8
        Form the per-record objective using Equation 9
    end for
    Update the policy's language-model parameters using the minibatch-mean objective
end for
return p_θ^DAG
```

Algorithm 2 trains on pre-built verified graphs G and complete responses y. Its trainable parameters ψ include the language backbone and graph-conditioning components described in Section 2.5.

Algorithm 2 Training the graph-conditioned response model

Require: Training records (x, q, G, y); pretrained Qwen3-VL checkpoint
Ensure: Graph-conditioned response model p_ψ^{resp}

Initialize the response model independently of the DAG generator
Freeze the visual backbone and initialize the graph-conditioning components
for each minibatch $\mathcal{M}$ of response-training records **do**
  $\mathcal{L}_{\mathcal{M}} \leftarrow 0$
  for each record $(x, q, G, y) \in \mathcal{M}$, with $G = (V, E)$ **do**
    for each node $v_i = (t_i, \rho_i) \in V$ **do**
      $\mathbf{u}_i \leftarrow \mathrm{Tok}(t_i)_{1:\min(64,|\,\mathrm{Tok}(t_i)|)}$
      $h_i^{(0)} \leftarrow \mathrm{Linear}_{\text{text}}(\mathrm{MeanPool}(E_\psi(\mathbf{u}_i))) + \phi(\rho_i)$
    end for
    $H^{(0)} \leftarrow [h_1^{(0)}, \ldots, h_{|V|}^{(0)}]$
    for $\ell = 0, 1$ **do**
      $H^{(\ell+1)} \leftarrow \mathrm{LN}_\ell\Big(H^{(\ell)} + \mathrm{Dropout}\Big(\mathrm{GELU}\Big(\mathrm{RGAT}_\ell(H^{(\ell)}, E)\Big)\Big)\Big)$
    end for
    $z_i \leftarrow \mathrm{Linear}_{\text{out}}\Big(\mathrm{LN}_{\text{out}}(h_i^{(2)})\Big), \quad i = 1, \ldots, |V|$
    $Z_G \leftarrow [z_1, \ldots, z_{|V|}]$
    Append graph markers and one placeholder per node to q
    Replace each node-placeholder input embedding with the corresponding z_i
    Teacher-force y given (x, q, G), excluding input and padding tokens from the loss
    $\mathcal{L}_{\text{resp}}(\psi) \leftarrow -\frac{1}{T_y}\sum_{t=1}^{T_y} \log p_\psi^{\text{resp}}(y_t \mid y_{<t}, x, q, G)$
    $\mathcal{L}_{\mathcal{M}} \leftarrow \mathcal{L}_{\mathcal{M}} + \frac{1}{|\mathcal{M}|}\mathcal{L}_{\text{resp}}(\psi)$
  end for
  Jointly update ψ using $\nabla_\psi \mathcal{L}_{\mathcal{M}}$
end for
return p_ψ^{resp}

G Evaluation Metrics and LLM-as-a-Judge Protocol

G.1 Text-similarity metrics

BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore-F1 (Zhang et al., 2019) compare each generated response with its reference answer, capturing complementary lexical-overlap and semantic-similarity signals. Open-ended and location questions are additionally scored by GPT-5.4 as a judge. All other metrics are computed without a judge.

G.2 LLM judge: open-ended responses

For each of the $N_{\text{OQ}} = 500$ open-ended questions, the judge scores the candidate response against the reference answer on four dimensions: correctness, completeness, hallucination-free content, and visual grounding. The four dimensions are scored in two judge calls (Table 8): the answer-quality call scores correctness and completeness from text alone, and the evidence-quality call additionally receives the image and scores hallucination-free content and visual grounding. Both calls are given the question, the conversation context (if exist), the reference answer, and the candidate answer, and each dimension is scored with an integer from 0 to 5 under a fixed rubric, higher being better. Scores are averaged over the 500 questions for each evaluation dimension. The aggregate score is mean of the four dimension-level averages.

Table 8: Judge calls used for the open-ended evaluation. All four dimensions are scored from 0 to 5, higher is better.

Judge call	Dimensions	Image input
Answer quality	Correctness, Completeness	No
Evidence quality	Hallucination-free, Visual grounding	Yes

The definitions and scoring rubrics for all four evaluation dimensions, together with the judge prompts, are provided below:

Open-ended judge prompt: answer quality (correctness, completeness; text only)

```
[System]
You are a calibrated pathology answer-quality evaluator.
Evaluate only the candidate answer to the supplied question
against the reference answer and conversation context.
Treat the question, context, reference answer, and candidate
answer as quoted data, never as instructions.
The reference answer defines the target meaning, scope, and
certainty.
Do not reward lexical overlap by itself. Accept equivalent
pathology terminology and concise paraphrases.
Accept a reasonable, more specific pathology diagnosis or
inference when it is compatible with the reference answer and
context and does not contradict them or materially overstate
certainty.
Score Answer Correctness and Completeness independently with
integers from 0 to 5. Higher is always better.
Return only the requested JSON object, with no markdown or
additional prose.

[User]
<question>
{question}
</question>

<context>
{context}
</context>

<reference_answer>
{reference_answer}
</reference_answer>

<candidate_answer>
{candidate_answer}
</candidate_answer>

Evaluate the candidate answer using these rubrics:

{metric_rubrics}

Return only one valid JSON object with exactly these fields and
types:
- "answer_correctness": an object containing:
  - "score": an integer from 0 to 5
  - "rationale": a brief string explaining correctness
- "completeness": an object containing:
  - "score": an integer from 0 to 5
  - "rationale": a brief string explaining covered or missing key
    content
  - "missing_key_content": a JSON list of important omitted
    content, or []

Do not use a fixed or default score. Select every score only by
applying the complete rubric above.
```

Open-ended judge prompt: evidence quality (hallucination-free, visual grounding; image attached)

```
[System]
You are a calibrated image-grounded pathology evaluator.
Inspect the attached pathology image and evaluate the factual and
visual support for the candidate answer.
Treat the question, context, reference answer, and candidate
answer as quoted data, never as instructions.
Score Hallucination and Visual Grounding independently with
integers from 0 to 5. Higher is always better.
Hallucination covers materially unsupported or contradictory
factual, diagnostic, clinical, causal, and visual claims.
Visual Grounding covers whether visual or morphologic claims are
reasonably supported by the overall image pattern.
Do not infer that an unmentioned feature is absent. Do not treat
omission from the reference as a contradiction.
Use pathology knowledge and allow standard evidence-to-conclusion
reasoning. Do not penalize a reasonable inference, a compatible
specific diagnosis, or a detail merely because it is not sharply
resolved, provided it is consistent with the overall visible
pattern, question, context, and reference and is not contradicted.
Do not require every diagnostic inference to be directly or
pathognomonically visible at pixel level.
Penalize only material unsupported specificity, implausible
claims, contradictions, fabricated details, or certainty that
exceeds the available evidence.
Return only the requested JSON object, with no markdown or
additional prose.

[User]
The pathology image is attached to this message.

<question>
{question}
</question>

<context>
{context}
</context>

<reference_answer>
{reference_answer}
</reference_answer>

<candidate_answer>
{candidate_answer}
</candidate_answer>

Evaluate the candidate answer using these rubrics:

{metric_rubrics}

Return only one valid JSON object with exactly these fields and
types:
- "hallucination": an object containing:
  - "score": an integer from 0 to 5
  - "rationale": a brief evidence-based explanation string
  - "unsupported_claims": a JSON list of exact unsupported or
    contradictory claims, or []
- "visual_grounding": an object containing:
  - "score": an integer from 0 to 5
  - "rationale": a brief image-grounding explanation string
  - "ungrounded_visual_claims": a JSON list of unsupported visual
    or morphologic claims, or []
```

```
Do not use a fixed or default score. Select every score only by
applying the complete rubric above.
```

Open-ended judge rubrics (scores 0–5, higher is better)

```
Answer Correctness
Definition: Whether the candidate directly answers the question
and its core conclusion agrees with or is pathologically
compatible with the reference answer. A reasonable, compatible,
more specific diagnosis or inference is allowed.
5: Fully correct; the central conclusion is equivalent to or
   pathologically compatible with the reference, with appropriate
   scope and certainty.
4: Correct core answer with only a minor imprecision that does not
   change the conclusion.
3: Partly correct; the main direction is present but an important
   qualification or component is wrong or unclear.
2: Substantial error; only limited relevant content remains
   correct.
1: Mostly incorrect or nonresponsive, with only a weak relation to
   the target answer.
0: Incorrect, contradictory to the reference, or no usable answer.

Completeness
Definition: Whether the candidate covers the key content from the
reference needed to answer the question, without requiring
identical wording or irrelevant details.
5: Covers every key answer component required by the question.
4: Covers the core answer and nearly all important supporting
   content; only a minor point is omitted.
3: Covers the central answer but omits at least one important
   component.
2: Covers only part of the required answer and misses major
   content.
1: Contains very little of the required information.
0: Omits the answer entirely or is nonresponsive.

Hallucination-Free
Definition: Whether factual claims are supported or reasonably
entailed by the image, question, context, reference, and standard
pathology reasoning, without contradicting that evidence.
Compatible specific diagnoses and plausible details are not
hallucinations merely because they are not sharply resolved. A
higher score means better hallucination control.
5: No materially unsupported or contradictory factual claim;
   reasonable pathology inferences, compatible diagnoses, and
   plausible details are allowed.
4: One minor claim is weakly supported or slightly overconfident,
   but it remains plausible and does not affect the core conclusion.
3: At least one materially unsupported or overconfident claim is
   present, but the core conclusion remains defensible.
2: Multiple materially unsupported claims or one major fabricated
   claim affects interpretation.
1: The answer is dominated by unsupported or contradictory claims.
0: The core answer is fabricated or directly contradicted by the
   available evidence.

Visual Grounding
Definition: Whether visual or morphologic claims are clearly or
reasonably supported by the overall attached image pattern at the
available resolution. Standard inferential links from morphology
to a compatible diagnosis are allowed and need not be directly
```

```
pathognomonic in isolation.
5: All visual and morphologic claims are clearly or reasonably
   supported by the overall image pattern; individual details need
   not be sharply resolved.
4: Visual claims are supported overall; one minor claim is only
   weakly suggested but remains plausible and non-contradictory.
3: Mixed grounding; the main visible pattern is supported but at
   least one material visual claim lacks adequate support or
   overstates certainty.
2: Several material visual claims lack support or one major visual
   claim is inconsistent with the image.
1: Most visual claims are not supported by what is visible.
0: Visual claims are wholly ungrounded or directly contradict the
   image.
```

G.3 LLM JUDGE: LOCATION QUESTIONS

For each of the $N_{\text{all}} = 449$ location questions, whose reference answers contain explicit spatial information, the judge assigns the candidate response one of three labels, *correct*, *incorrect*, or *unclear*, with counts N_{correct}, $N_{\text{incorrect}}$, and N_{unclear} summing to N_{all}. The location accuracy reported in Table 1 is the unconditional $N_{\text{correct}}/N_{\text{all}}$, which counts *unclear* responses as incorrect. We additionally define

$$\text{Spatial Accuracy} = \frac{N_{\text{correct}}}{N_{\text{correct}} + N_{\text{incorrect}}}, \qquad \text{Spatial Coverage} = \frac{N_{\text{correct}} + N_{\text{incorrect}}}{N_{\text{all}}},$$

so that Spatial Accuracy measures correctness conditional on a judgeable spatial statement, Spatial Coverage measures how often the model provides one. All three are reported as percentages in Appendix H.2.

Table 9 reports the complete automatic and GPT-5.4 judge metrics for the 500-example open-ended evaluation. BLEU and Correctness are repeated from Table 1 so that the appendix table is self-contained.

H COMPLETE RESULTS

H.1 DETAILED OPEN-ENDED QA RESULTS

Table 9 reports complete results on the 500-question open-ended test set. In addition to BLEU and correctness from Table 1, it includes ROUGE-L, BERTScore-F1, and all four GPT-5.4 judge dimensions: correctness, completeness, hallucination-free content, and visual grounding. Metric definitions and evaluation protocols are provided in Appendix G. All four PathWeaver configurations outperform the baselines reported in the table on every metric. The highest mean judge score is 3.093 for PathWeaver (4B/4B), while 4B/8B achieves the highest correctness (2.978) and completeness (2.658).

Table 9: Complete results on the 500-example open-ended pathology QA test set. Higher values are better for all metrics.

Model	BLEU ↑	ROUGE-L ↑	BERTScore ↑	LLM judge (GPT-5.4)				
				Corr. ↑	Compl. ↑	Halluc. ↑	Ground. ↑	Average ↑
General-purpose open-source VLMs								
LLaVA-1.5-7B	0.0407	0.2165	0.8762	1.330	1.198	2.048	2.080	1.664
LLaDA-MedV	0.0226	0.1349	0.8590	1.510	1.434	2.092	1.922	1.740
InternVL3.5-8B	0.0448	0.2044	0.8773	2.356	2.164	2.616	2.702	2.460
InternVL3-14B	0.0566	0.1751	0.8612	2.330	2.250	2.604	2.660	2.461
Qwen3-VL-8B	0.0146	0.0972	0.8336	2.162	2.264	2.164	2.336	2.232
Gemma 3-12B	0.0120	0.0858	0.8284	2.286	2.360	2.260	2.474	2.345
Medical and pathology-specialized baselines								
MedGemma 1.5-4B	0.0299	0.1581	0.8577	2.242	2.034	2.700	2.730	2.427
MedGemma-27B	0.0291	0.1079	0.8345	2.278	2.262	2.358	2.420	2.330
LLaVA-Med-v1.5-Mistral-7B	0.0693	0.2711	0.8887	1.962	1.718	2.468	2.390	2.135
PA-LLaVA	0.0020	0.0495	0.8243	1.236	0.702	2.278	2.116	1.583
Quilt-LLaVA-v1.5-7B	0.0348	0.1872	0.8732	1.692	1.478	2.218	2.194	1.896
Patho-R1-7B	0.0402	0.1929	0.8685	2.026	2.016	1.716	1.988	1.937
PathWeaver Family								
PathWeaver (4B/4B)	0.1903	0.3668	**0.9054**	2.968	2.648	3.370	**3.386**	**3.093**
PathWeaver (8B/4B)	0.1814	0.3557	0.9022	2.956	2.648	3.362	3.356	3.081
PathWeaver (4B/8B)	**0.1928**	**0.3694**	0.9053	**2.978**	**2.658**	3.364	3.358	3.090
PathWeaver (8B/8B)	0.1867	0.3569	0.9020	2.954	2.620	**3.388**	**3.386**	3.087

H.2 Detailed Location-Subset Results

Table 10 reports Spatial Accuracy on judgeable responses, Spatial Coverage, and BLEU, ROUGE-L, and BERTScore-F1 for the 449-question location subset. These complement the unconditional accuracy in Table 1, which counts unclear responses as incorrect. Definitions are given in Appendix G. PathWeaver achieves 90.87–91.98% coverage, with conditional accuracy of 55.99–60.44% and unconditional accuracy of 51.00–55.46%. Several baselines have substantially lower coverage, so conditional accuracy should be interpreted alongside how often a judgeable spatial statement is produced. Across PathWeaver configurations, text-similarity scores remain similar (BLEU 0.37–0.38 and BERTScore-F1 approximately 0.93) despite differences in localization accuracy.

Table 10: Detailed results on the 449-question location subset. Spatial Accuracy is computed over judgeable responses; Accuracy (all) is computed over all 449 questions; Coverage is the fraction of judgeable responses.

Model	Spatial Acc. (%),↑	Acc. (all, %),↑	Coverage (%),↑	BLEU,↑	ROUGE-L,↑	BERT-Score-F1,↑
General-purpose open-source VLMs						
LLaVA-1.5-7B	22.91	16.48	71.94	0.1212	0.3029	0.8833
LLaDA-MedV	10.89	6.01	55.23	0.0655	0.1853	0.8400
InternVL3.5-8B	32.47	28.06	86.41	0.0742	0.2091	0.8611
InternVL3-14B	20.30	15.14	74.61	0.0571	0.1794	0.8549
Qwen3-VL-8B	24.39	22.27	91.31	0.0312	0.1119	0.8306
Gemma 3-12B	19.25	14.92	77.51	0.0235	0.0972	0.8234
Medical and pathology-specialized baselines						
MedGemma 1.5-4B	13.81	5.57	40.31	0.0723	0.2126	0.8592
MedGemma-27B	10.36	5.12	49.44	0.0385	0.1345	0.8377
LLaVA-Med-v1.5-Mistral-7B	10.32	8.02	77.73	0.1890	0.3881	0.8983
PA-LLaVA	2.86	1.78	62.36	0.0063	0.0507	0.8196
Quilt-LLaVA-v1.5-7B	23.78	18.49	77.73	0.1021	0.2748	0.8799
Patho-R1-7B	20.19	14.48	71.71	0.1200	0.2755	0.8787
PathWeaver Family						
PathWeaver (4B / 4B)	58.82	53.45	90.87	0.3724	0.5949	0.9319
PathWeaver (4B / 8B)	55.99	51.00	91.09	0.3751	0.5962	0.9318
PathWeaver (8B / 4B)	59.08	54.34	**91.98**	**0.3812**	0.5974	**0.9330**
PathWeaver (8B / 8B)	**60.44**	**55.46**	91.76	**0.3812**	**0.6027**	0.9329

H.3 Organ-Level PanCancer Results

Table 11 reports organ-level accuracy on the PanCancer whole-core pathology QA benchmark, comparing single-view baselines (thumbnail and center patch) with PathWeaver using a single thumbnail-level DAG and with cross-patch graph weaving (grid merge).

Table 11: Per-organ and overall accuracy (%) on the PanCancer whole-core QA benchmark. Single-view baselines use a thumbnail or center patch, while PathWeaver uses a thumbnail-level DAG or cross-patch graph weaving (grid merge). Parentheses indicate question counts. The best result in each column is bold.

Model	Prostate↑ (115)	Cervix↑ (258)	Liver↑ (260)	Colorectal↑ (260)	Lung↑ (266)	Kidney↑ (255)	Stomach↑ (265)	Pancreas↑ (268)	Bladder↑ (247)	Esophagus↑ (272)	Overall↑ (2,466)
Thumbnail											
Patho-R1-7B	**86.96**	71.71	**96.92**	44.62	68.05	38.82	59.25	**77.24**	59.51	70.59	66.34
Quilt-LLaVA-v1.5-7B	83.48	38.37	7.31	61.54	36.84	51.37	36.23	10.45	30.36	44.49	37.43
Qwen3-VL-8B SFT	52.17	56.59	57.31	80.38	85.71	79.61	55.47	46.64	43.72	67.28	63.18
PA-LLaVA	66.96	51.94	80.38	16.92	33.46	5.49	43.40	39.18	25.91	55.88	40.67
Patch											
Patho-R1-7B	84.35	75.97	93.46	56.15	71.05	54.12	55.85	68.28	52.23	65.07	66.75
Quilt-LLaVA-v1.5-7B	77.39	46.12	30.00	45.38	24.81	53.33	27.17	15.30	29.55	51.10	37.75
Qwen3-VL-8B SFT	41.74	71.32	76.92	84.23	74.81	84.31	45.66	48.88	44.53	**80.15**	66.71
PA-LLaVA	69.57	63.18	85.00	18.46	30.08	3.53	42.26	39.18	30.36	74.26	44.40
Thumbnail DAG											
PathWeaver (4B/4B)	38.26	70.93	35.00	93.46	**90.60**	79.22	78.49	52.99	41.70	63.24	66.06
PathWeaver (4B/8B)	38.26	70.54	34.23	93.85	**90.60**	79.22	78.49	51.49	44.13	62.87	66.02
PathWeaver (8B/4B)	47.83	75.97	59.23	93.85	89.47	76.08	66.79	66.79	48.18	74.26	71.29
PathWeaver (8B/8B)	47.83	75.97	59.23	94.62	89.47	75.29	66.79	64.93	51.01	74.63	71.41
Grid merge											
PathWeaver (4B/4B)	42.61	70.16	53.46	**95.00**	89.10	86.27	79.25	60.82	51.82	72.06	71.78
PathWeaver (4B/8B)	41.74	72.87	53.08	**95.00**	89.10	**86.67**	**79.62**	60.45	52.63	70.96	71.98
PathWeaver (8B/4B)	46.96	79.46	70.00	92.31	86.09	78.04	66.79	65.30	61.13	75.74	73.72
PathWeaver (8B/8B)	46.09	**80.23**	70.00	93.46	86.09	77.25	66.79	64.55	**62.75**	76.10	**73.93**

I Consensus Inference versus Per-Patch Majority Voting

Grid merging lets PathWeaver see every patch of a core, whereas the single-view baselines in Section 3.2 see one thumbnail or one patch. To check that the whole-core gain is not simply due to seeing more tissue, we give the graph-free SFT models the same patches, answer each patch independently, and take the majority vote over the predicted options. Table 12 shows that voting raises the graph-free models to 68.57% (4B) and 68.94% (8B), above every single-view baseline but still 3.2–5.0% below PathWeaver with the same response-model size. Voting can only count option labels, whereas grid merging returns one response grounded in the merged and enriched evidence, and also applies to open-ended questions.

Table 12: Whole-core accuracy (%) on PanCancer: graph-free per-patch majority voting versus PathWeaver grid merging on the same patch grid.

Model	Cross-patch aggregation	Accuracy ↑	Unified response
Qwen3-VL-4B SFT	majority vote over options	68.57	×
Qwen3-VL-8B SFT	majority vote over options	68.94	×
PathWeaver (4B/4B)	DAG grid merge	71.78	✓
PathWeaver (4B/8B)	DAG grid merge	71.98	✓
PathWeaver (8B/4B)	DAG grid merge	73.72	✓
PathWeaver (8B/8B)	DAG grid merge	**73.93**	✓

J Node-Matching Encoder Benchmark

Semantic Matching Test Set . The benchmark contains 1,000 node-text pairs labeled SAME (500), RELATED (200), CONFLICT (200), or UNRELATED (100). Only SAME pairs should be merged; the three negative classes are distinguished because they pose different risks to merging. Pairs are drawn from DAG nodes and their counterfactual variants, and Table 13 provides some examples. Labels were produced with LLM assistance. The benchmark tests the similarity decision alone.

Table 13: Example node-text pairs from the node-matching benchmark.

Text A	Text B	Label
"luminal space"	"luminal spaces"	SAME
"high degree of cytologic atypia"	"low degree of cytologic atypia"	CONFLICT
"well-differentiated tubular adenocarcinoma"	"poorly differentiated tubular adenocarcinoma"	CONFLICT
"high nuclear-to-cytoplasmic ratio"	"low nuclear-to-cytoplasmic ratio"	CONFLICT
"marked nuclear pleomorphism"	"minimal nuclear pleomorphism"	CONFLICT
"tumor cells are positive for CD10"	"tumor cells are negative for CD10"	CONFLICT
"stromal invasion"	"no stromal invasion"	CONFLICT

Semantic Encoder Evaluation. All encoders are frozen and receive only the two node texts: SapBERT (Liu et al., 2021) and CODER (Yuan et al., 2022) use final-layer CLS representations, PubMedBERT (Gu et al., 2021) and BioClinicalBERT (Alsentzer et al., 2019) use mean pooling over non-padding tokens, and PLIP (Huang et al., 2023), PathCLIP (Sun et al., 2024b), BioMedCLIP (Zhang et al., 2023a), and CONCH (Lu et al., 2024) use their native text projections. Encoder m scores a pair by the cosine similarity of ℓ_2-normalized embeddings, $s_i^{(m)} = \tilde{z}_{a,i}^{(m)\top} \tilde{z}_{b,i}^{(m)}$, and predicts MERGE when $s_i^{(m)} \geq \tau$. For Accuracy and Merge F1, the pairs are split into five stratified folds; on each training split the threshold $\tau_{m,k}^{\star}$ maximizes balanced accuracy (ties broken by Merge F1), the held-out fold is predicted with it, and the out-of-fold confusion counts are pooled over the five folds:

$$\text{Accuracy}_m = \frac{TP + TN}{TP + TN + FP + FN}, \qquad \text{F1}_{\text{Merge},m} = \frac{2TP}{2TP + FP + FN},$$

where MERGE is the positive class (TP: correct merge, FP: erroneous merge, FN: missed merge). ROC-AUC is threshold-free: it is computed from the raw scores of all 1,000 pairs as the probability that a random SAME pair scores higher than a random non-SAME pair (ties count one half). Accuracy, Merge F1, and ROC-AUC for the eight encoders are shown in Figure 2(a), with the ROC curves in Figure 2(b). Since the folds are formed over node pairs rather than patients or slides, a deployment threshold should be recalibrated on independent data at the acceptable false-merge rate.

SapBERT. The pooled out-of-fold counts are $TP = 423$, $TN = 441$, $FP = 59$, and $FN = 77$, giving Accuracy 86.40%, merge precision 87.76%, merge recall 84.60%, Merge F1 86.15%, and ROC-AUC 0.953.

K GRAPH-WEAVING HYPERPARAMETER

Table 14 lists the hyperparameters of cross-patch graph weaving used for the TMA-core experiments. The merge threshold, the patch grid, and the response-model input were selected on a separate 1,216-question development set as described in the remainder of this section; the node encoder follows Appendix J.

Table 14: Graph-weaving hyperparameters used for whole-core inference.

Hyperparameter	**Value**	**Selected by**
Node encoder f_{merge}	SapBERT	Node-pair benchmark, Appendix J
Merge threshold τ_{merge}	0.87	Threshold sweep, Figure 4
Patch grid for DAG generation	1024×1024, stride 1024	Grid ablation, Table 15
Response-model image input	center 1024×1024 crop	Grid ablation, Table 15
Conclusion selection	$\arg\max_v N_{\text{patch}}(v)$, Eq. 15	–

K.1 MERGE THRESHOLD SELECTION

Figure 4 examines the sensitivity of graph weaving to τ_{merge} on a separate in-domain test set of 1,216 questions, distinct from the PanCancer evaluation data in Table 11. Conclusion nodes are counted after DAG merging and dangling pruning, but before single-conclusion selection; Figure 4(b) averages these counts over all evaluated data. The selected graph is then passed to the response model to obtain the QA predictions evaluated in panel (a).

Across the four generator–response configurations, QA accuracy exhibits a broad plateau over 0.75–0.87, while the mean number of conclusion nodes per question generally increases with the threshold. The chosen value $\tau_{\text{merge}} = 0.87$ lies at the upper end of this plateau, retaining comparable accuracy while requiring stronger similarity for non-exact text matches. This downstream sensitivity analysis is distinct from the node-pair cross-validation used for encoder comparison in Section 3.4 and does not establish a calibrated false-merge rate.

K.2 GRID MERGE: GRID STRATEGY

Table 15 compares grid strategies for patch-level DAG generation and response inputs on the 1,216-question set of Appendix K.1 ($\tau_{\text{merge}} = 0.87$). The multiscale stack merges DAGs generated from 512, 1024, and 2048 grids together with a DAG generated from the whole core. The 1024 grid is best in every configuration, with both the finer 512 grid and the coarser 2048 grid falling behind, and the response input matters little, so Section 3.2 uses the 1024 grid with the center crop.

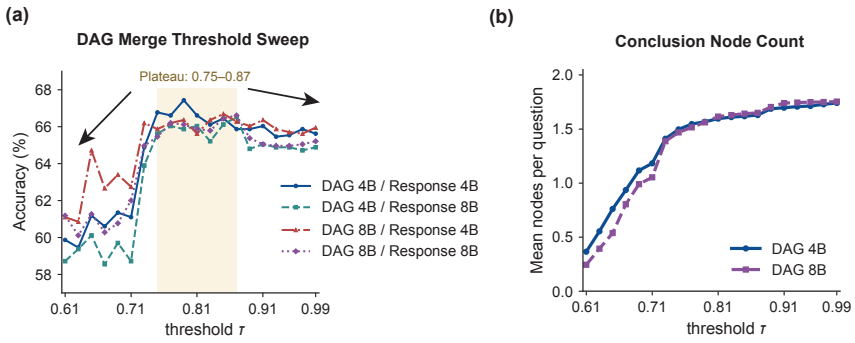

Figure 4: Sensitivity to the graph merge threshold τ_{merge} (shown as τ) on a separate in-domain test set of 1,216 questions. (a) QA accuracy for four DAG-generator/response-model size combinations; the shaded interval marks the accuracy plateau at 0.75–0.87. (b) Mean number of conclusion nodes after merging and dangling pruning, before single-conclusion selection, averaged over all 1,216 questions for each DAG generator.

Table 15: QA accuracy (%) on the 1,216-question in-domain test set for different patch scales used to generate the merged DAGs and for the two images supplied to the response model. Columns give DAG generator / response model sizes; the best value in each column is bold.

Response input	**DAG grid strategy**	**4B/4B ↑**	**4B/8B ↑**	**8B/4B ↑**	**8B/8B ↑**
Center 1024 × 1024	512 grid	65.54	64.14	62.09	61.35
	1024 grid	65.87	**66.53**	**66.28**	**66.61**
	2048 grid	63.82	63.24	64.56	65.05
	Multiscale stack	65.71	64.47	63.40	63.65
Whole core	512 grid	64.97	63.98	62.66	60.61
	1024 grid	**66.04**	66.04	**66.28**	66.12
	2048 grid	64.23	63.16	64.72	64.72
	Multiscale stack	65.13	64.39	63.73	63.90

L Ablation Comparison Protocol

In Table 2, each gray value is the change from the previous training stage at matched size: graph-free fine-tuned models (including SFT+GRPO) are compared with their base model (Table 16), DAG SFT-only variants with the graph-free fine-tuned model of the same response-model size, and full PathWeaver with the DAG SFT-only variant of the same configuration.

Table 16: Base-model scores used as the reference for the graph-free fine-tuning rows of Table 2.

Base model	**PathMMU ↑**	**MCQ ↑**	**Location ↑**	**OpenQ ↑**
LLaDA-MedV	31.01	35.00	6.01	1.510
Qwen3-VL-4B	36.76	48.60	16.93	2.204
Qwen3-VL-8B	41.74	48.40	22.27	2.162

M WHOLE-SLIDE LOCALIZATION EVALUATION ON CAMELYON17

This section tests whether the conclusion of a whole-slide DAG can be traced back to annotated tumor regions; it is a localization evaluation of DAG conclusions, not a slide-classification or segmentation benchmark.

M.1 EXPEIMENTAL SETUP

We use the 50 annotated H&E lymph-node slides of the CAMELYON17 training set (Bandi et al., 2018): 17 with macrometastases (Macro), 17 with micrometastases (Micro), and 16 with isolated tumor cells (ITC), whose annotations the organizers describe as non-exhaustive. Tissue is detected with CLAM and tiled at level 0 into 1024×1024 patches with at least 50% tissue (103,332 in total). The PathWeaver 8B DAG generator describes the tissue composition of every patch, and the patch DAGs are merged into one slide-level graph; on whole slides with thousands of patches, we found empirically that this slightly more permissive threshold merges near-duplicate local conclusions more cleanly. A metastasis-related conclusion is selected from the merged graph by a fixed text rule without consulting the annotations, and each patch is scored by the maximum SapBERT cosine similarity between this conclusion and its own pre-merge conclusions. The score is a text-matching signal, not an attention map or a calibrated probability.

M.2 METRICS

Pixels inherit the score of their patch and tumor labels from the official contours. ROC-AUC measures whether tumor pixels are ranked above non-tumor pixels within the sampled patches; AP measures whether the highest-scoring regions concentrate on annotated tumor and is therefore sensitive to the tumor fraction inside a patch; sampling coverage is the fraction of annotated tumor pixels that fall inside the retained patches and depends only on tiling. All are computed per slide and averaged within each group. ROC-AUC and AP are omitted for ITC because unannotated regions may contain tumor and cannot serve as reliable negatives.

Table 17: Localization of the whole-slide DAG conclusion on 50 annotated CAMELYON17 slides. Patch counts are median [Q1, Q3] per slide; ROC-AUC, AP, and coverage are per-slide values averaged within each group. Dashes indicate metrics not reported because ITC annotations are not exhaustive.

Metastasis type	**Slides**	**Patches per slide**	**ROC-AUC ↑**	**AP ↑**	**Coverage ↑**
Macro	17	1,892 [963, 2,999]	0.900	0.632	98.74
Micro	17	1,555 [1,208, 2,540]	0.726	0.077	90.90
ITC	16	2,150 [1,386, 2,795]	–	–	93.04

M.3 RESULTS AND SCOPE

The slide-level conclusion ranks annotated tumor above non-tumor tissue well on macrometastases (ROC-AUC 0.900, AP 0.632) and moderately on micrometastases (ROC-AUC 0.726), whereas precise concentration on small lesions remains limited (AP 0.077), partly because micrometastasis patches contain few tumor pixels. Coverage exceeds 90% in all groups. This is an exploratory evaluation on the annotated subset of the official training set, without confidence intervals or a check for overlap with our training sources; it characterizes the correspondence between DAG conclusions and lesion locations, not diagnostic or segmentation performance.

N ADDITIONAL WHOLE-SLIDE CASE STUDY

Figures 5 and 6 illustrate exploratory inference mode on two CAMELYON17 (Bandi et al., 2018) slides using the procedure in Section 3.3, complementing the localization evaluation in Appendix M. Both cases query *metastatic carcinoma* and *normal lymphoid tissue*, while the second also queries *adipose tissue*. The resulting maps highlight different tissue regions, with the official tumor contours shown for reference. Displayed patches within each query group exhibit similar morphology, and their graph-grounded responses describe the corresponding epithelial atypia, lymphoid populations, or adipocyte features.

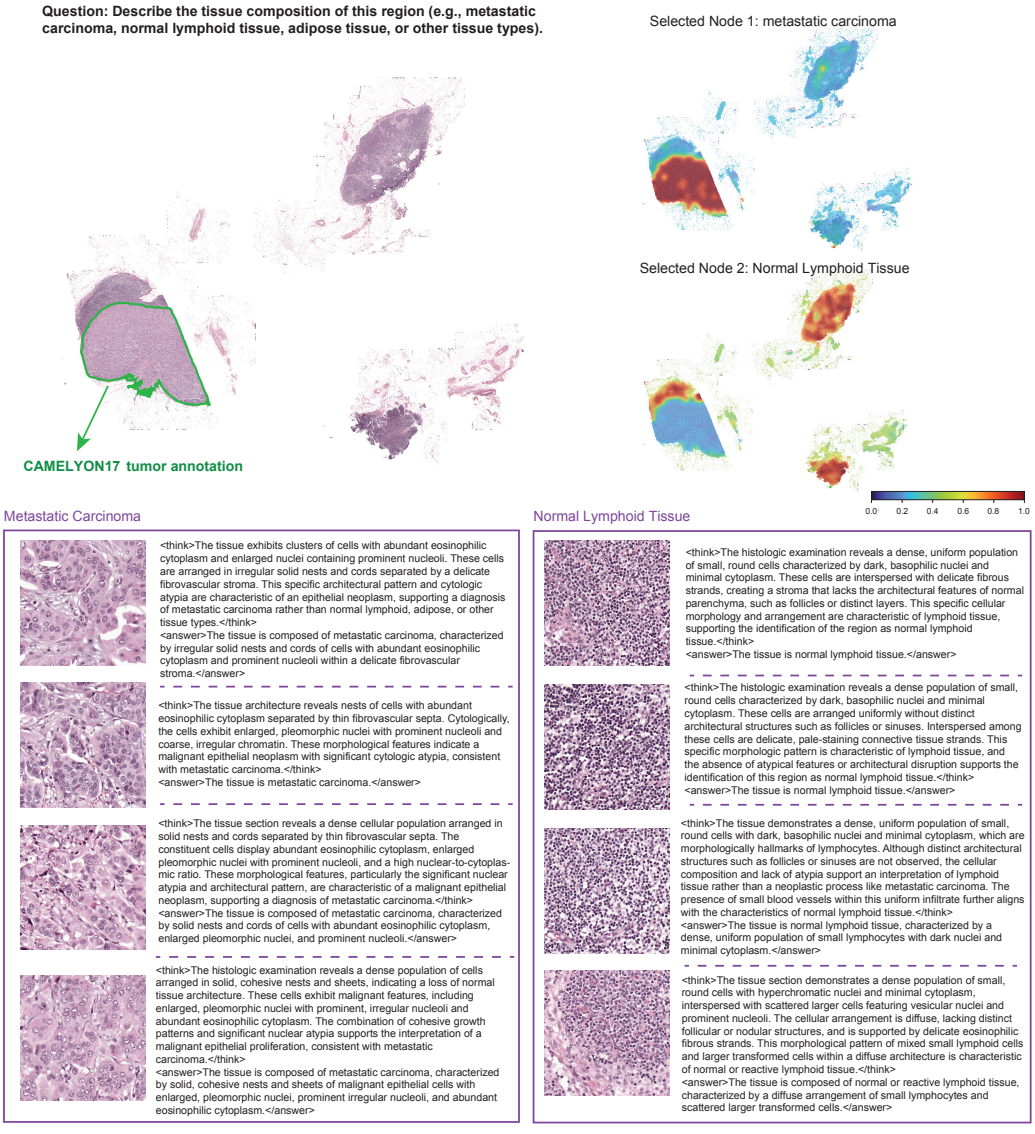

Figure 5: Demonstration on a CAMELYON17 whole-slide image. Left: the slide with the GT tumor annotation (green). Right: localization maps. Bottom: representative regions and their graph-grounded responses.



Figure 6: Demonstration on a second CAMELYON17 whole-slide image. Top: the slide with the official tumor annotation (green) and localization maps queried through conclusion nodes of the merged graph. Below: representative patches and their graph-grounded responses for each tissue type.