PATHMEM: TOWARD COGNITION-ALIGNED MEMORY TRANSFORMATION FOR PATHOLOGY MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Computational pathology demands both visual pattern recognition and dynamic integration of structured domain knowledge, including taxonomy, grading criteria, and clinical evidence. In practice, diagnostic reasoning requires linking morphological evidence with formal diagnostic and grading criteria. Although multimodal large language models (MLLMs) demonstrate strong vision–language reasoning capabilities, they lack explicit mechanisms for structured knowledge integration and more reliable control. As a result, existing models struggle to consistently incorporate pathology-specific diagnostic standards during reasoning. Inspired by the hierarchical memory process of human pathologists, we propose `PathMem`, a memory-centric multimodal framework for pathology MLLMs. `PathMem` organizes structured pathology knowledge as a long-term memory (LTM) and introduces a **Memory Transformer** that models the dynamic transition from LTM to working memory (WM) through multimodal memory activation and context-aware knowledge grounding, enabling context-aware memory refinement for downstream reasoning. `PathMem` achieves SOTA performance across benchmarks, improving WSI-Bench report generation (**+13.0%** WSI-P, **+11.1%** WSI-R) and open-ended diagnosis by **+9.7%** and **+8.9%** over prior WSI-based models. Code: PathMem.

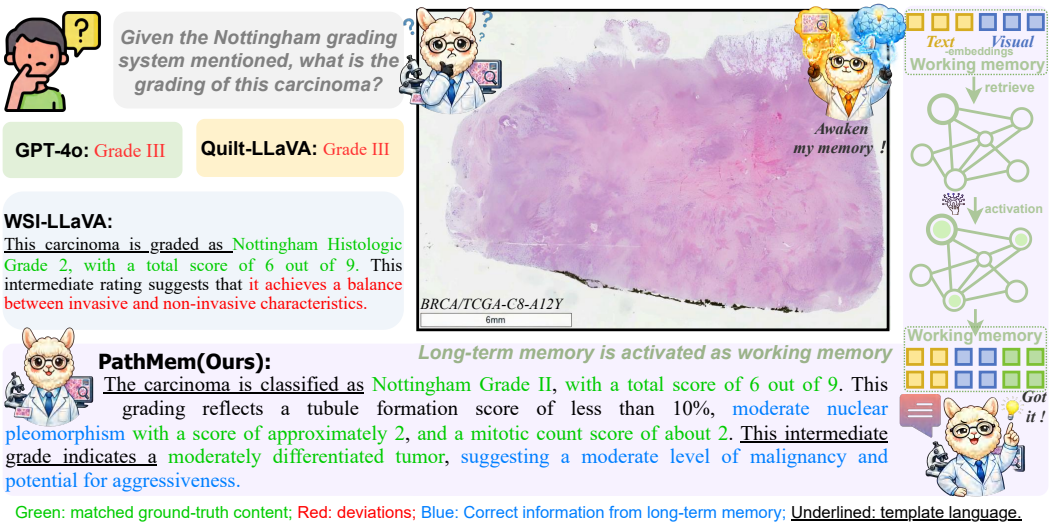

Figure 1: Overview of `PathMem`. The working memory selectively activates long-term memory traces and transforms them into an updated working memory state for more reliable reasoning.

1 INTRODUCTION

Computational pathology is a knowledge-intensive and cognitively complex discipline (Omar et al., 2024; Liu et al., 2026; Cifci et al., 2023; Liang et al., 2025; Sun et al., 2025; Ahmed et al., 2024; Lu

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.

et al., 2024; Chen et al., 2025a). Accurate diagnosis requires not only visual pattern recognition from histopathological images but also the dynamic integration of long-term expert knowledge (Verghese et al., 2023), including disease taxonomy (Shafique et al., 2024; Redline et al., 2023; Romero et al., 2022; Gabitto et al., 2024), grading criteria (Jiang et al., 2024; Gosai et al., 2025; Li et al., 2025), and evolving clinical evidence (Fountzilas et al., 2025; Ullah et al., 2025; Bian et al., 2026). In practice, diagnostic reasoning requires linking morphological observations with formal diagnostic and grading standards. Although recent multimodal large language models (MLLMs) (Yin et al., 2024; Caffagni et al., 2024; Liu et al., 2023) have demonstrated promising vision-language reasoning capabilities, they largely operate as parametric black boxes and lack explicit mechanisms for structured knowledge integration and more reliable memory control (Liu et al., 2025a; Cabalo et al., 2024; Sheng et al., 2025). Therefore, incorporating structured, controllable, and more reliable knowledge integration mechanisms into multimodal reasoning architectures may become a key challenge for building the next generation of clinically reliable and cognitively robust intelligent systems in computational pathology.

In clinical practice, pathologists rely on a hierarchical memory process (Hou et al., 2024; Hu et al., 2025; Liu et al., 2025b; Chen et al., 2025b; Sun et al., 2024a): accumulated domain knowledge forms long-term memory (LTM) (Bei et al., 2026; Yang et al., 2022; Minhas et al., 2024; Blackmore et al., 2023), while case-specific evidence selectively activates relevant knowledge into working memory (WM) for reasoning (Baddeley, 2025; Akhtar et al., 2024; Karakatsani et al., 2023; Manchanda et al., 2023), and iteratively refine their hypotheses. Existing retrieval-augmented generation (RAG) methods attempt to introduce external knowledge (Zhang et al., 2025; Hamza et al., 2025; Jabal et al., 2025; Omar et al., 2024; Fink et al., 2025), but their retrieval pipelines are typically static and fail to model the dynamic and adaptive nature of human memory transformation. This cognitive process, characterized by memory retrieval, selective activation, and meta-cognitive regulation, is largely absent in current MLLMs (An, 2025; Huang et al., 2026; Jiang et al., 2026a; Hou et al., 2025; Yoshida et al., 2025; Dong et al., 2025).

In this paper, we propose `PathMem`, a memory-centric multimodal architecture for pathology MLLMs (Figure 1). We first construct a **high-quality pathology knowledge graph** as LTM via deep semantic search over PubMed, augmented with diverse sources, including clinical guidelines, gray literature, among others. The knowledge graph is encoded into semantic embeddings aligned with the multimodal backbone. We then design a **Memory Transformer** that explicitly models the transition from LTM to WM. Given visual and textual embeddings, we perform: Static activation, which ranks knowledge entries via cosine similarity; Dynamic activation, which jointly projects multimodal and knowledge embeddings to compute global relevance. An adaptive selection strategy determines the boundary of activated knowledge, transferring only highly relevant entries into WM for final reasoning. This mechanism enables controlled and more reliable memory enhancement during inference. In report generation and open-ended diagnosis, it improves WSI-P / WSI-R by **13.0% / 11.1%** and **9.7% / 8.9%**, respectively.

The contributions of this paper are summarized as follows.

- ❶ **Memory-Driven Pathology MLLM Architecture**: We propose `PathMem`, an explicit long-term/working memory paradigm into multimodal pathology architecture, enabling cognitively grounded and knowledge-aware reasoning.
- ❷ **High-Quality LTM Construction**: We build a structured pathology knowledge graph via deep search over PubMed, providing a scalable and updatable repository that simulates accumulated expert-level domain knowledge.
- ❸ **Dynamic–Static Memory Controller with Adaptive Activation**: We propose a dual-mode memory activation mechanism with self-adaptive selection, explicitly modeling the transformation from LTM to WM for context-aware, more reliable, and enhanced inference.
- ❹ **PathMem achieves the SOTA performance across multiple Benchmarks**: In report generation, it improves WSI-P by **13.0%** and WSI-R by **11.1%**. In open-ended diagnosis, it further yields gains of **9.7%** and **8.9%**, respectively.

💡 **By replacing static retrieval with explicit memory transformation, our framework enables more reliable, memory-aware reasoning aligned with human cognition in pathology MLLMs.**

2 Related Work

Pathology MLLM. With the rapid development of computational pathology, the availability of large-scale whole slide images (WSIs) paired with diagnostic reports has catalyzed the rise of pathology-oriented MLLMs (Xu et al., 2024; Ding et al., 2025; Ahmed et al., 2024; Chen et al., 2024; Allaberdiev et al., 2026; Chen et al., 2025a). For instance, Prov-GigaPath (Xu et al., 2024) and TITAN (Ding et al., 2025) emphasize large-scale whole-slide modeling and vision–language alignment, while PRISM (Shaikovski et al., 2024) and PathAlign (Ahmed et al., 2024) demonstrate the effectiveness of language-aligned slide embeddings for zero-shot and generative applications. Building upon these slide-level foundations, recent pathology MLLMs further integrate instruction tuning, task unification, and computational efficiency into a cohesive framework. PathAsst (Sun et al., 2024b) and CPath-Omni (Sun et al., 2025) construct unified multimodal assistants that consolidate patch- and slide-level tasks under large-scale image–text supervision. SlideChat (Chen et al., 2025d) and WSI-LLaVA (Liang et al., 2025) enhance morphology-aware reasoning and conversational capabilities through dedicated WSI-benchmarks and instruction data. To address the prohibitive cost of gigapixel modeling, LoC-Path (Hu et al., 2025) introduces redundancy-aware compression for efficient slide representation, while ChatEXAONEPath (Kim et al., 2025) strengthens expert-level clinical reasoning via retrieval-based data construction. ♣ In summary, while *prior pathology MLLMs achieve strong slide-level understanding through large-scale pretraining and alignment, they lack explicit mechanisms to balance long-term knowledge and short-term contextual memory, limiting stable and consistent clinical reasoning.*

Knowledge in pathology. Explicit domain knowledge is essential for robust computational pathology (Ma et al., 2025; Smart & Roux, 1995; Huang et al., 2025; Jiang et al., 2026b; Zhao et al., 2024). Knowledge has been incorporated as an inductive bias during pretraining and adaptation. Unified knowledge distillation aggregates expert and self-knowledge to enhance cross-task generalization (Ma et al., 2025), while knowledge-guided adaptation leverages visual and textual priors to improve cross-domain robustness and demographic fairness (Huang et al., 2025). Knowledge concept-based MIL frameworks connect slide representations with disease-specific linguistic concepts derived from medical literature (Zhao et al., 2024), enabling semantically grounded modeling. In multimodal prognosis prediction, prior and textual knowledge guide phenotype extraction and genotype interaction (He et al., 2025; Shi et al., 2025), reducing modality gaps and enhancing clinically meaningful integration. Despite progress, most knowledge-enhanced methods rely on implicit supervision or proprietary RAG pipelines, limiting reproducibility and long-term knowledge utilization. Different from **prior knowledge-injection** strategies that treat knowledge as auxiliary supervision or external retrieval, *we approach enhancement from a **memory-aware perspective** by constructing a high-quality pathology knowledge base and integrating it as structured LTM within pathology MLLMs, enabling persistent, explicit, and more reliable reasoning.*

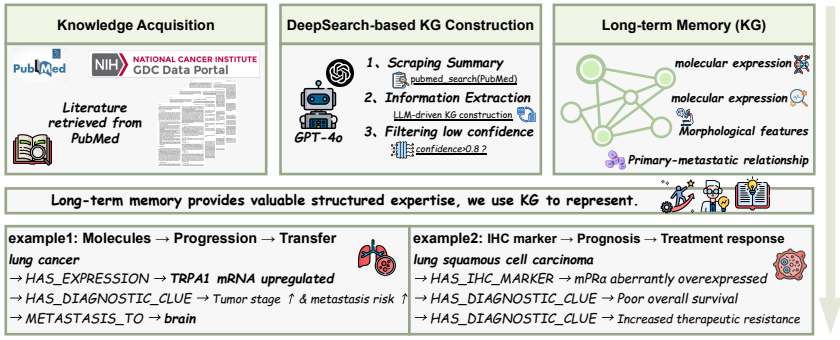

Figure 2: LTM construction pipeline for pathology knowledge graphs via iterative literature retrieval and LLM-based information extraction.

3 METHOD

3.1 UNIFIED MODEL INTERFACE

Problem Formulation. We formulate LTM construction as a literature-grounded knowledge aggregation problem. Given a domain-specific query q_0, we construct a weighted directed multigraph $\mathcal{G}$ encoding disease–feature–evidence relations extracted from biomedical abstracts. Formally, the resulting LTM is defined as

$$\mathcal{G} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{E}}, \mathbf{W}, \Phi, \Psi)\,, \tag{1}$$

where $\mathcal{V}$ denotes the global entity set, $\boldsymbol{\mathcal{R}}$ denotes the predefined relation schema, $\boldsymbol{\mathcal{E}} \subseteq \boldsymbol{\mathcal{V}} \times \boldsymbol{\mathcal{R}} \times \boldsymbol{\mathcal{V}}$ denotes the set of directed relational triples, $\mathbf{W} : \boldsymbol{\mathcal{E}} \to [0, 1]$ assigns calibrated confidence weights to edges, Φ denotes the entity normalization mapping, and Ψ denotes the feature-based inverted index for retrieval. The objective is to maximize structured knowledge coverage while controlling redundancy and uncertainty propagation.

3.2 LTM CONSTRUCTION

As illustrated in Figure 2, we construct the LTM through an iterative pipeline that integrates literature retrieval with LLM-based information extraction to progressively build and refine the pathology knowledge graph. The statistics of the constructed pathology knowledge graph are presented in Table 1. ♣ **Hash-Based Deduplication and Memory Monotonicity.** To guarantee monotonic growth of memory without redundancy, we define a deterministic hashing operator $\boldsymbol{\mathcal{H}}(\cdot)$ and compute $h_i = \boldsymbol{\mathcal{H}}\left(\mathrm{Normalize}(a_i)\right)$, where $\mathrm{Normalize}(\cdot)$ removes formatting noise, punctuation variation, and case differences before hashing. Let $\boldsymbol{\mathcal{M}}_h^{(t)}$ denote the set of previously observed hash values at iteration t. The deduplicated abstract set is then defined as

$$\boldsymbol{\mathcal{A}}^* = \left\{ a_i \;\middle|\; h_i \notin \boldsymbol{\mathcal{M}}_h^{(t)} \right\}, \quad \boldsymbol{\mathcal{M}}_h^{(t+1)} = \boldsymbol{\mathcal{M}}_h^{(t)} \cup \{h_i\}\,, \tag{2}$$

here h_i is the hash signature of abstract a_i, $\boldsymbol{\mathcal{M}}_h^{(t)}$ is the memory of previously seen abstracts, and $\boldsymbol{\mathcal{A}}^*$ is the effective corpus after deduplication. This guarantees that identical textual evidence does not repeatedly influence the long-term graph. ❥ **LLM-Based Structured Knowledge Extraction.** Each deduplicated abstract $a_i \in \boldsymbol{\mathcal{A}}^*$ is processed by a large language model extractor f_θ parameterized by θ: $\mathcal{T}_i = f_\theta(a_i) = \{(s_{ij}, r_{ij}, o_{ij}, c_{ij}, \mathbf{z}_{ij})\}_{j=1}^{K_i}$, here K_i is the number of candidate triples extracted from abstract a_i, $s_{ij} \in \boldsymbol{\mathcal{V}}$ is the subject entity, $r_{ij} \in \boldsymbol{\mathcal{R}}$ is the relation type, $o_{ij} \in \boldsymbol{\mathcal{V}}$ is the object entity, $c_{ij} \in [0, 1]$ is the model-estimated confidence score, and $\mathbf{z}_{ij} \in \mathbb{R}^d$ is the latent semantic embedding of the triple, capturing contextual representation for downstream similarity consolidation. ✤ **Confidence-Constrained Triple Filtering.** To ensure high precision in LTM, we introduce a minimum confidence threshold τ. The retained triple set from abstract a_i is defined as

$$\boldsymbol{\mathcal{T}}_i^{\tau} = \{(s_{ij}, r_{ij}, o_{ij}, c_{ij}, \mathbf{z}_{ij}) \in \boldsymbol{\mathcal{T}}_i \mid c_{ij} \geq \tau\}\,, \tag{3}$$

here $\tau \in (0, 1)$ controls the trade-off between recall and precision, and is empirically determined on a validation set to balance precision and coverage. Only triples whose estimated confidence exceeds τ are injected into the knowledge graph. ♠ **Probabilistic Multi-Evidence Fusion.** Since identical semantic triples may appear across multiple abstracts, we perform probabilistic evidence aggregation. For a canonical triple (s, r, o) appearing in m independent sources, with confidence scores $\{c^{(1)}, c^{(2)}, \ldots, c^{(m)}\}$, the aggregated edge weight is defined as

$$\boldsymbol{\mathcal{W}}(s, r, o) = 1 - \prod_{k=1}^{m} \left(1 - \alpha \cdot c^{(k)} \cdot \exp\left(-\boldsymbol{\mathcal{F}} \left\| \mathbf{z}^{(k)} - \bar{\mathbf{z}} \right\|_2^2\right)\right), \tag{4}$$

where $\alpha \in (0, 1]$ is a global scaling coefficient, $c^{(k)}$ is the confidence score from the k-th abstract, $\mathbf{z}^{(k)}$ is the embedding representation of the k-th evidence instance, $\bar{\mathbf{z}}$ is the centroid embedding of all supporting instances, $\|\cdot\|_2$ denotes Euclidean norm, and $\boldsymbol{\mathcal{F}} > 0$ controls embedding-consistency penalization. This formulation follows a noisy-or probabilistic fusion mechanism, where multi-source corroboration increases edge certainty, while embedding inconsistency reduces effective contribution. ♦ **Feature-Oriented Memory Indexing.** To support efficient retrieval during reasoning, we construct a feature-inverted index $\boldsymbol{\Psi}(f) = \{(s, r, o) \in \boldsymbol{\mathcal{E}} \mid f = s \;\vee\; f = o\}$, where $f \in \boldsymbol{\mathcal{V}}$ denotes a histopathological feature entity, and $\Psi(f)$ returns all triples involving that feature. Additionally,

disease nodes are defined as $\mathcal{V}_{\text{disease}} = \{v \in \mathcal{V} \mid \text{Type}(v) = \text{Disease}\}$. These memory indices enable subgraph extraction from microscopic features. Literature retrieval is performed on PubMed using entity-augmented keyword queries with BM25 ranking (top-$K = 50$ per query), and structured extraction is conducted by a pretrained GPT-4-class LLM ($T = 0$) with a fixed prompt template.

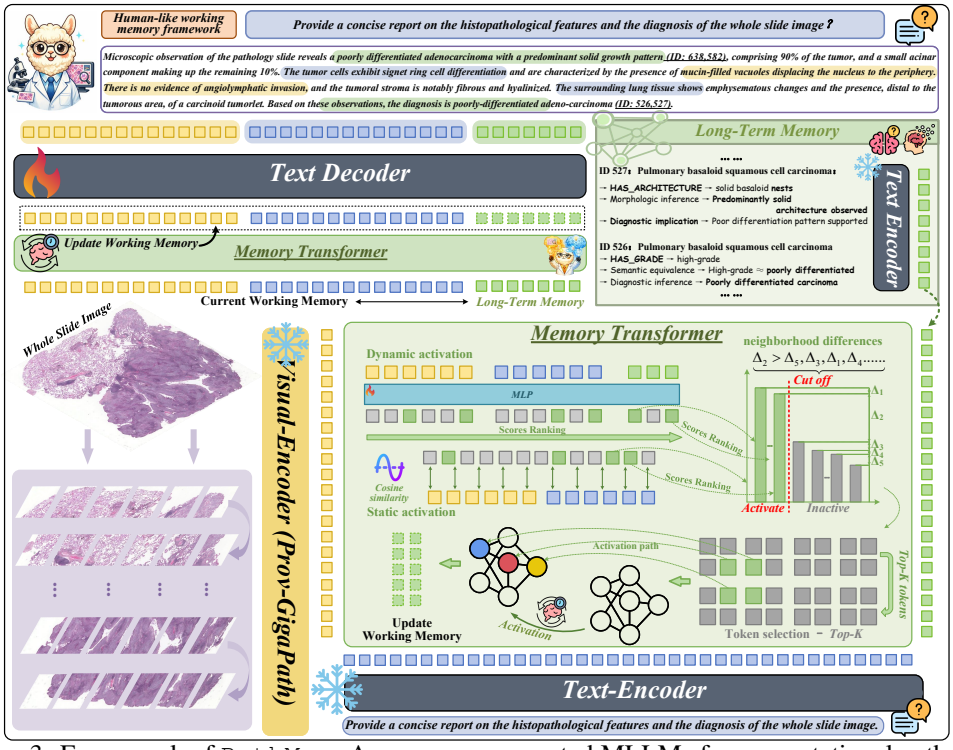

Figure 3: Framework of PathMem. A memory-augmented MLLMs for computational pathology that aligns visual, textual, and knowledge graph representations, and adaptively activates LTM for knowledge-grounded reasoning about pathology.

Table 1: Statistics of the pathology knowledge graph.

Category	Metric	Value
Graph size	Unique nodes	9,974
	Unique directed edges	10,748
	Relation types	19
Source coverage	Accepted pathology articles	1,998
	Unique PMIDs in the initial pool	2,914
Connectivity	Largest component	93.9%
	Approximate diameter	11
	Avg. path length	4.78
Disease complexity	Disease nodes	490
	Mean feature facts per disease	22.36
	Mean relation types per disease	5.63
Expert supervision	Pathologists	4
	Participating centers	3
	Countries	2

♣ **Pathology Knowledge Graph Statistics and Validation.** The construction and validation of the pathology knowledge graph were ***supervised by four pathologists from three medical centers across two countries.*** The pathologists performed article-level screening to determine the literature included in the knowledge base and reviewed each constructed KG to assess the validity and pathological plausibility of the extracted knowledge and relations.

3.3 Memory Transformer: Knowledge-Aware LTM→WM Transition

As illustrated in Figure 3, the **Memory Transformer** serves as the core module that bridges the LTM and WM through structured embedding interactions. ♣ **LTM Embedding Space.** Let the constructed knowledge graph define a structured LTM space represented by an embedding bank

$$\mathcal{Q} = \left\{ \mathbf{m}_i \in \mathbb{R}^d \,\middle|\, i = 1, \ldots, N \right\}, \qquad \mathcal{Q} \in \mathbb{R}^{N \times d}, \tag{5}$$

here N denotes the total number of knowledge entities stored in LTM, d is the shared embedding dimensionality with the multimodal backbone, and each $\mathbf{m}_i$ encodes a graph-aware knowledge representation learned from literature-grounded triples. The memory space $\mathcal{Q}$ is fixed during inference and serves as an external, persistent knowledge reservoir. Given an input multimodal sequence $\mathbf{X} \in \mathbb{R}^{T \times d}$, we compute a normalized query representation $\mathbf{q} = \bar{\mathbf{x}}_t \left(\|\bar{\mathbf{x}}_t\|_2 + \varepsilon \right)^{-1}$, here $\mathbf{x}_t$ denotes the t-th token embedding, $\| \cdot \|_2$ is the Euclidean norm, and ε is a small constant for

numerical stability. The vector $\mathbf{q} \in \mathbb{R}^d$ summarizes the global semantic context of the current input. ➽ **Structured Memory Selection.** We define a relevance operator over the LTM space:

$$\mathcal{J} = \text{Softmax}\left(\frac{\mathcal{Q}\mathbf{q}}{\sqrt{d}} + \mathcal{N}\right), \qquad \mathcal{J} \in \mathbb{R}^N. \tag{6}$$

In this formulation, $\mathcal{Q}\mathbf{q}$ denotes the similarity between all memory tokens and the query, $\frac{1}{\sqrt{d}}$ is the scaled dot-product normalization factor, and $\mathcal{N} \in \mathbb{R}^N$ is a structural masking vector that suppresses previously selected indices or enforces diversity constraints, we adopt a non-replacement top-k masking strategy, where previously selected indices are assigned $-\infty$ in $\mathcal{N}$ to prevent repeated selection and encourage diversity. The resulting vector $\mathcal{J}$ represents a probability distribution over the entire memory bank, encoding query-conditioned relevance scores. To obtain a sparse WM, we apply top-k selection:

$$\mathcal{I} = \text{TopK}\left(\mathcal{J}, k\right), \quad \mathcal{Q}_{\text{work}} = \mathcal{J}_{\mathcal{I}} \odot \mathcal{Q}_{\mathcal{I}}, \tag{7}$$

here $k \ll N$ ensures sparsity, $\mathcal{I}$ denotes the selected indices, $\mathcal{Q}_{\mathcal{I}} \in \mathbb{R}^{k\times d}$ is the retrieved subset, and $\odot$ denotes row-wise scaling by relevance weights. This operation converts dense LTM into a compact WM block. ♠ **Memory-Augmented Transformer Input.** The WM is then prepended to the original sequence: $\mathbf{X}^\star = \text{Concat}\left(\mathcal{Q}_{\text{work}}, \mathbf{X}\right) \in \mathbb{R}^{(k+T)\times d}$. The augmented sequence $\mathbf{X}^\star$ is fed into the Transformer encoder, allowing self-attention to jointly model long-term knowledge tokens and input-specific representations. Through this mechanism, external structured memory is dynamically projected into the model's space, enabling evidence-aware reasoning without expanding parameters. The model supports incremental memory updates without requiring full retraining of the Memory Transformer.

4 Experiments

❶ **Experimental Setup.** Experiments are primarily conducted on **WSI-Bench**, with additional zero-shot evaluation on external datasets to assess generalization. More details on data preprocessing are provided in Appendix B.1. ❷ **In-domain benchmark.** We adopt **WSI-Bench** as the main benchmark for training and evaluation. Derived from TCGA, it contains 9,850 WSIs spanning 30 tumor types and 179,569 VQA pairs, covering morphology understanding, diagnosis, therapy reasoning, and report generation. We follow the official split, using 9,642 slides (175,450 QA pairs) for training and 208 slides (4,119 QA pairs) for testing. Evaluation follows the original protocol, including WSI-P, WSI-R, accuracy, BLEU, and ROUGE. ❸ **Out-of-domain evaluation.** We further assess generalization in a zero-shot manner on three external datasets: *WSI-VQA* (Chen et al., 2025c), *SlideBench-VQA (BCNB)* (Chen et al., 2025d), and *CPTAC-NSCLC*. These datasets are used strictly for evaluation without additional fine-tuning. Except for Tables 2 and 3, the default LLM used is Llama. The details of **Long-term memory construction** are included in Appendix A. ❹ **Implementation Details & Metrics** are included in Appendix B.2.

4.1 Main Result

Obs. ❶: Consistent Performance Gains of `PathMem` Across Pathological Tasks. Table 2 presents the comparison on WSI-Bench. `PathMem` (LLaMA) achieves an overall average of 0.768, surpassing WSI-LLaVA (Liang et al., 2025) at 0.754, Quilt LLaVA at 0.721, WSI-VQA (Chen et al., 2025c) at 0.590, and GPT-4o (Islam & Moushi, 2024) at 0.507, demonstrating superior comprehensive capability. In Morphological Analysis, `PathMem` (LLaMA) attains competitive open ended precision and close ended accuracy at 0.535 and 0.972, with relevance at 0.542. For Diagnosis, it leads in precision and relevance with 0.707 and 0.701, while achieving 0.868 accuracy. In Treatment Planning, `PathMem` (LLaMA) achieves 0.796 precision, 0.792 relevance, and 1.000 accuracy, while the Qwen3 variant also shows consistently competitive performance across the three tasks.

Table 2: Quantitative evaluation of WSI and general-purpose MLLMs on our WSI-Bench across three pathological capabilities. WSI-P: WSI-Precision, WSI-R: WSI-Recall, Acc: accuracy, open: open-ended question, and close: close-ended question.

Model	Input	Morphological Analysis			Average	Diagnosis			Average	Treatment Planning			Average
		Open		Close		Open		Close		Open		Close	
		WSI-P	WSI-R	Acc		WSI-P	WSI-R	Acc		WSI-P	WSI-R	Acc	
Quilt-LLaVA	Thumbnail	0.448	0.447	0.947	0.614	0.586	0.604	0.849	0.679	0.788	0.816	**1.000**	0.868
GPT-4o	Thumbnail	0.220	0.204	0.471	0.298	0.472	0.457	0.530	0.486	0.496	0.841	0.875	0.737
WSI-VQA	WSI	0.395	0.462	0.758	0.538	0.436	0.525	0.577	0.512	0.791	0.595	0.771	0.719
WSI-LLaVA	WSI	0.488	0.610	0.951	0.683	0.610	0.612	0.863	0.695	**0.810**	**0.845**	**1.000**	**0.885**
`PathMem (LLaMA)`	WSI	0.535	0.542	**0.972**	0.683	**0.707**	**0.701**	0.868	**0.758**	0.796	0.792	**1.000**	0.863
`PathMem (Qwen3)`	WSI	**0.598**	**0.623**	0.960	**0.727**	0.640	0.639	**0.882**	0.720	0.796	0.796	0.979	0.857

Table 3: Evaluation in report generation on WSI-Bench.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	WSI-P	WSI-R
GPT-4o	0.202	0.069	0.030	0.016	0.132	0.167	0.067	0.138
Quilt-LLaVA	0.474	0.351	0.282	0.236	0.475	0.460	0.324	0.333
WSICaption	0.403	0.306	0.248	0.209	0.446	0.407	0.310	0.377
TITAN	0.399	0.321	0.223	0.203	0.304	0.385	—	—
Hist-Gen	0.406	0.307	0.248	0.208	0.448	0.416	0.300	0.397
WSI-LLaVA	0.480	0.358	0.287	0.240	0.490	0.465	0.380	0.429
PathAgent	0.505	0.368	0.313	0.257	0.501	0.488	0.433	0.449
SlideChat	0.441	0.310	0.277	0.191	0.463	0.422	0.331	0.340
MLLM-HWSI	0.556	0.426	0.348	0.287	0.551	0.513	0.502	0.518
PathMem (LLaMA)	**0.561**	**0.433**	**0.360**	**0.311**	**0.556**	**0.530**	**0.510**	**0.540**
PathMem (Qwen3)	0.548	0.427	0.341	0.295	0.539	0.524	0.492	0.538

Obs. ❷: Superior Report Generation Performance of PathMem. Table 3 reports the results for report generation. PathMem (LLaMA) achieves the best performance across all metrics, with BLEU-1 to BLEU-4 scores of 0.561, 0.433, 0.360, and 0.311, consistently outperforming WSI-LLaVA (Liang et al., 2025) (0.480, 0.358, 0.287, 0.240) and other baselines. In semantic evaluation, PathMem (LLaMA) attains 0.556 ROUGE-L and 0.530 METEOR, surpassing WSI-LLaVA (Liang et al., 2025) (0.490 / 0.465) and significantly exceeding GPT-4o (Islam & Moushi, 2024) (0.132 / 0.167), indicating stronger content coverage and alignment with reference reports. It also achieves the highest WSI-P and WSI-R (0.510 / 0.540), compared with 0.380 / 0.429 for WSI-LLaVA (Liang et al., 2025), reflecting improved slide-level grounding and clinical relevance. Overall, consistent gains across lexical, semantic, and WSI-specific metrics, with the Qwen3 variant showing similarly competitive performance, demonstrate the superiority of PathMem in generating accurate and pathology-coherent reports. **Obs. ❷: Superior Report Generation Performance of PathMem.** Table 3 reports the results for report generation. PathMem (LLaMA) achieves the best performance across all metrics, with BLEU-1 to BLEU-4 scores of 0.561, 0.433, 0.360, and 0.311, consistently outperforming WSI-LLaVA (Liang et al., 2025) (0.480, 0.358, 0.287, 0.240) and other baselines. In semantic evaluation, PathMem (LLaMA) attains 0.556 ROUGE-L and 0.530 METEOR, surpassing WSI-LLaVA (Liang et al., 2025) (0.490 / 0.465) and significantly exceeding GPT-4o (Islam & Moushi, 2024) (0.132 / 0.167), indicating stronger content coverage and alignment with reference reports. It also achieves the highest WSI-P and WSI-R (0.510 / 0.540), compared with 0.380 / 0.429 for WSI-LLaVA (Liang et al., 2025), reflecting improved slide-level grounding and clinical relevance. Overall, consistent gains across lexical, semantic, and WSI-specific metrics, together with the competitive results of the Qwen3 variant, demonstrate the superiority of PathMem in generating accurate and pathology-coherent reports.

Table 4: Zero-shot validation on external benchmarks.

Model	Input	WSI-VQA	SlideBench-VQA (BCNB)				CPTAC
			Tumor	Grade	Subtype	Avg	
Quilt- LLaVA	Thumb	0.132	0.113	0.194	0.181	0.163	0.603
WSI-VQA	WSI	0.469	0.039	0.300	0.000	0.113	0.000
SlideChat	WSI	-	0.902	0.231	0.175	0.436	-
WSI-LLaVA	WSI	0.546	0.905	0.463	0.292	0.553	0.721
PathMem (Ours)	WSI	**0.572**	**0.918**	**0.482**	**0.313**	**0.571**	**0.754**

Obs. ❸: Strong Zero-Shot Generalization of PathMem Across External Benchmarks. Table 4 presents the zero-shot external validation results. PathMem achieves the best performance on all benchmarks, demonstrating strong generalization across datasets and tasks. On WSI-VQA (Chen et al., 2025c), PathMem attains 0.572, outperforming WSI-LLaVA (Liang et al., 2025) at 0.546 and substantially exceeding Quilt-LLaVA (Seyfioglu et al., 2024) at 0.132. On SlideBench-VQA BCNB, PathMem achieves the highest scores in all categories, including Tumor at 0.918, Grade at 0.482, and Subtype at 0.313, yielding the best average of 0.571. In comparison, WSI-LLaVA (Liang et al., 2025) obtains an average of 0.553, while WSI-VQA (Chen et al., 2025c) shows clear instability with an average of 0.113 and zero performance on Subtype. On CPTAC NSCLC, PathMem reaches 0.754, surpassing WSI-LLaVA (Liang et al., 2025) at 0.721 and markedly improving over Quilt-LLaVA (Seyfioglu et al., 2024) at 0.603.

Obs. ❹: The Superior Reasoning Capability of PathMem. As shown in Figure 4, we present a qualitative comparison for case **LUSC/TCGA-21-1083**. The ground truth is *poorly differentiated (Grade III), non-keratinizing squamous cell carcinoma*. **Overall Diagnosis.** PathMem correctly predicts poorly differentiated squamous cell carcinoma, fully consistent with the reference. WSI-LLaVA (Liang et al., 2025) captures squamous features but outputs an inconsistent hybrid diagnosis. GPT-4o (Islam & Moushi, 2024) and Quilt-LLaVA (Seyfioglu et al., 2024) misclassify the case as adenocarcinoma variants, reflecting confusion between squamous and glandular differentiation. **Fine-Grained Morphology.** For squamous differentiation (Claims 1, 3–5), PathMem best captures poor differentiation and marked pleomorphism, with partial recognition of non-keratinizing features. Other models show weaker alignment and often hallucinate glandular structures. For nuclear atypia (Claim 2), all models detect pleomorphism, but PathMem demonstrates greater consistency. For necrosis and invasion (Claims 6–7), performance is generally limited; however, PathMem correctly identifies the absence of vascular invasion. For background lung changes (Claim 8), WSI-LLaVA (Liang et al., 2025) and GPT-4o (Islam & Moushi, 2024) note emphysematous features, whereas PathMem focuses on tumor characteristics.

Obs. ❺: Metric Evidence Supporting Reasoning Strength of `PathMem`. The quantitative results closely mirror the color-coded qualitative comparison (*Appendix Figure 6*). **Hist-Gen** (Guo et al., 2024) exhibits numerous red deviations and orange omissions, leading to low semantic overlap (BLEU-4: 0.025; ROUGE-L: 0.182) and limited pathological relevance (WSI-P: 0.257; WSI-R: 0.230). **WSI-LLaVA** (Liang et al., 2025) shows moderate improvement, with fewer deviations but continued omission of key invasive features. This corresponds to slightly higher BLEU-4 (0.043), ROUGE-L (0.250), and WSI-based metrics (WSI-P: 0.311; WSI-R: 0.294). In contrast, `PathMem` demonstrates predominantly green matches with minimal deviations, alongside multiple blue knowledge graph–grounded concepts. This results in substantially improved lexical and clinical alignment (BLEU-4: 0.361; ROUGE-L: 0.583; METEOR: 0.678) and markedly higher WSI-P and recall (0.596 / 0.592). Overall, reductions in deviations and missing findings correspond directly to gains across both NLU and WSI-based metrics, while knowledge graph grounding further enhances diagnostic consistency. **♣ Answer-Aligned Memory Path.** Visual Cues (solid growth, poor differentiation, minimal glandular formation, angiolymphatic invasion, emphysematous lung background) → **Static LTM Retrieval** (lung primary site, NSCLC spectrum, solid/high-grade architecture) → **Dynamic Multimodal Reweighting** (↑ adenocarcinoma-related morphology signals; ↓ irrelevant entities) → **Adaptive LTM→WM Transfer** (lung context + adenocarcinoma-related semantics + invasion knowledge) → **WM-Integrated Diagnosis: Invasive Poorly Differentiated Adenocarcinoma** (+ **angiolymphatic invasion**; + **focal visceral pleural involvement**; + **background emphysema**). The highlighted diagnostic elements appear in the generated report; the memory path above provides an illustrative interpretation of LTM→WM integration.



Figure 4: Qualitative comparison of generated reports by our method and three baseline approaches on the report generation task. (Red highlights denote incorrect content, while green highlights denote correct content, and orange highlights denote missing content from T-answer.)

Table 5: Ablation study on static and dynamic KG retrieval. The full model with both static and dynamic LTM achieves the best performance across all metrics. **Ret.** denotes retrieval.

Version	External Knowledge	Static Ret.	Dynamic Ret.	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	WSI-P	WSI-R
a	✓			0.481	0.357	0.287	0.241	0.477	0.464	0.404	0.445
b	✓	✓		0.528	0.405	0.332	0.284	0.518	0.508	0.492	0.516
c	✓		✓	0.521	0.396	0.324	0.278	0.510	0.511	0.489	0.510
d	✓	✓	✓	0.561	0.433	0.360	0.311	0.556	0.530	0.510	0.540

4.2 Analysis and Discussion

Obs. ❶: Ablation Study on the Effect of External Knowledge. Table 5 presents the ablation results on static and dynamic KG retrieval. Both mechanisms consistently improve performance over the baseline (Version a). Adding Static LTM (Version b) increases BLEU-4 from 0.241 to 0.284 and WSI-P from 0.404 to 0.492. dynamic LTM (Version c) also yields clear gains, with BLEU-4 reaching 0.278 and WSI-R 0.510.dynamic LTM The full model (Version d) achieves the best results on all metrics. These results indicate that static and dynamic retrieval provide complementary benefits, and their combination is essential for optimal performance. ***Attribution analysis of performance improvement:***

- **Gains from External Knowledge Integration.** Version b & c primarily reflect the benefit of incorporating external knowledge through static or dynamic retrieval mechanisms, respectively. Their improvements over the baseline can be largely attributed to the introduction of high-quality pathology knowledge.
- **Gains from Memory Control Mechanism** Version d introduces the full memory controller that jointly orchestrates dynamic and static knowledge activation. Therefore, the additional gains observed from Version d over Versions b and c can be more directly attributed to the effectiveness of the proposed memory control mechanism.

The memory controller plays a critical role in coordinating heterogeneous knowledge sources and enabling their complementary benefits, beyond what can be achieved by naive retrieval alone.

Table 6: Sensitivity analysis on the maximum number of activated KG tokens (Top-K). Performance comparison under different limits for dynamic and static LTM retrieval.

Version	Token Cap		PathMem							
	Dynamic-activation	Static-activation	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	WSI-P	WSI-R
a	1	1	0.517	0.393	0.320	0.274	0.508	0.500	0.477	0.496
b	2	2	0.525	0.402	0.330	0.284	0.517	0.509	0.484	0.497
c	3	3	0.531	0.410	0.339	0.293	0.525	0.514	0.493	0.513
d	4	4	0.537	0.414	0.340	0.292	0.528	0.517	0.502	0.526
g	5	5	**0.542**	**0.418**	**0.346**	**0.298**	**0.532**	**0.523**	**0.506**	**0.528**

Obs. ❷: Sensitivity Analysis of Top-K KG Tokens. Table 6 presents the sensitivity analysis on the maximum number of activated KG tokens for dynamic and static LTM retrieval. Performance generally improves, with a slight BLEU-4 decrease at K=4. With Top-K equals 1, BLEU-4 is 0.274 and WSI-P is 0.477. Increasing Top-K to 3 raises BLEU-4 to 0.293 and WSI-R to 0.513. The best results are achieved at Top-K equals 5, with 0.298 BLEU-4, 0.532 ROUGE-L, 0.523 METEOR, 0.506 WSI-P, and 0.528 WSI-R. The gains diminish beyond Top-K equals 3, for example BLEU-4 increases only from 0.293 to 0.298 when moving from 3 to 5. This indicates that a moderate token budget captures most useful knowledge.

Table 7: Inference efficiency analysis on WSI-Bench.

Metrics	Quilt-LLaVA	WSI-LLaVA	SlideChat	MLLM-HWSI	PathMem
Time (s/WSI) ↓	0.75	0.93	0.69	294	0.69
GPUs (GB) ↓	12.62	16.70	11.58	25.82	16.3

Table 8: Efficiency and practicality analysis of PathMem.

Metric	Value
Inference Time	∼1.45 items (WSI) / s
GPU Utilization	1 × GPU
GPU Memory Usage	16.3 GB
Image Feature Size	1728 × 4096 tokens
KG Construction Cost	Offline embedding + indexing
KG Update Cost	Re-embedding + re-indexing

Obs. ❸: Inference Efficiency Analysis. As listed in Table 7, PathMem achieves the lowest inference latency of 0.69 s/WSI (generation only, excluding WSI feature extraction), on par with SlideChat and substantially faster than MLLM-HWSI, while requiring 16.3 GB of GPU memory. This indicates that PathMem maintains high inference throughput with a moderate memory footprint. As listed in Table 8, PathMem runs at ∼1.45 items/s with 16.3 GB memory on a single GPU, indicating moderate inference overhead with good scalability. The additional cost mainly comes from KG construction and updates (embedding and indexing), which are performed offline and amortized over repeated inference. Overall, PathMem provides a practical efficiency–performance trade-off for high-quality reasoning tasks. Considering that computational pathology workflows are typically performed at slide-level or case-level rather than under strict real-time constraints, ***the current inference efficiency is sufficient for real-world clinical pathology deployment and routine diagnostic workflows.***

5 Conclusion & Limitation

We propose PathMem, a memory-augmented MLLM for computational pathology that explicitly models long- and working-memory through a PubMed-derived pathology knowledge graph and a dual-stage Memory Transformer. Experiments on WSI-Bench and three zero-shot benchmarks show consistent improvements over SOTA models across core pathology tasks. Ablation and sensitivity studies confirm the effectiveness of dual-mode KG retrieval and moderate Top-K token settings.

Limitation. PathMem introduces additional design complexity, which may require careful tuning of hyperparameters across tasks. The current implementation focuses on selected pathology benchmarks, and extending evaluation to a wider range of datasets could further assess generality. Moreover, the framework assumes a fixed knowledge retrieval strategy, leaving room for exploring more adaptive or task-specific retrieval mechanisms in future work.

LLM USAGE DISCLOSURE

Large Language Models (LLMs) were used as auxiliary tools in the preparation of this manuscript. In particular, GPT-5.6 supported language refinement, including grammar correction, sentence rephrasing, and improvements to clarity, coherence, and academic style. It was also used for limited assistance with general writing and coding-related tasks. All substantive research content, including the problem formulation, methodology, experimental design, analysis, and interpretation of results, was developed independently by the authors. LLM-generated suggestions were used only as supporting material and were carefully reviewed, verified, and revised before being incorporated into the manuscript. GPT-5.6 was not used to generate scientific findings, formulate conclusions, or make autonomous research decisions.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Section 3 details the PathMem architecture, including long-term memory construction, static and dynamic memory activation, and adaptive transfer to working memory. Appendix A and Algorithm 1 describe the literature-based knowledge graph construction pipeline, including evidence extraction, deduplication, confidence filtering, and memory indexing. Appendix B provides the dataset processing procedures, model configurations, training hyperparameters, and evaluation protocols, including additional experiments under a stricter split that excludes training slides sharing a case with the test set. Appendix C describes the claim-level evaluation procedure and scoring criteria, while Appendix D reports the computational setup.

ETHICS STATEMENT

PathMem is developed for research on knowledge-enhanced computational pathology using existing datasets and literature-derived knowledge. Although the framework improves performance on the evaluated benchmarks, these results do not establish clinical safety or readiness for deployment. Generated reports and diagnostic responses may contain unsupported findings, omit relevant evidence, or activate knowledge that is inappropriate for a particular case. Moreover, literature-derived memory may inherit publication biases, outdated evidence, and errors introduced during automated extraction; confidence filtering does not guarantee clinical validity. Dataset and pretrained-model biases may also lead to uneven performance across patient populations, institutions, and tumor types. PathMem should therefore not serve as a standalone basis for diagnosis or treatment decisions. Any clinical application would require independent validation, appropriate oversight, and review by qualified professionals. Subsequent use and distribution of data, models, and literature-derived resources should respect patient privacy and the applicable access and licensing conditions.

REFERENCES

Faruk Ahmed, Andrew Sellergren, Lin Yang, Shawn Xu, Boris Babenko, Abbi Ward, Niels Olson, Arash Mohtashamian, Yossi Matias, Greg S Corrado, et al. Pathalign: A vision-language model for whole slide images in histopathology. arXiv preprint arXiv:2406.19578, 2024.

Ansab Akhtar, Siddharth Singh, Ravinder Kaushik, Rajendra Awasthi, and Tapan Behl. Types of memory, dementia, alzheimer's disease, and their various pathological cascades as targets for potential pharmacological drugs. Ageing Research Reviews, 96:102289, 2024.

Sharofiddin Allaberdiev, Asif Khan, Sardor Mamarasulov, and Xiaojun Chen. Chestxgen: Dynamic memory-augmented vision-language transformer with context-aware gating for radiology report generation. Journal of Artificial Intelligence and Soft Computing Research, 16(1):55–72, 2026.

Tao An. Cognitive workspace: Active memory management for llms–an empirical study of functional infinite context. arXiv preprint arXiv:2508.13171, 2025.

Alan Baddeley. Working memory. In Memory, pp. 101–152. Routledge, 2025.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. arXiv preprint arXiv:2601.03515, 2026.

Jiang Bian, Yifan Peng, Eneida Mendonca, Imon Banerjee, and Hua Xu. Biomedical multimodal large language models: From model-centric development to clinically grounded evaluation and integration, 2026.

Daniel G Blackmore, Daniel Razansky, and Jürgen Götz. Ultrasound as a versatile tool for short-and long-term improvement and monitoring of brain function. Neuron, 111(8):1174–1190, 2023.

Donna Gift Cabalo, Jordan DeKraker, Jessica Royer, Ke Xie, Shahin Tavakol, Raúl Rodríguez-Cruces, Andrea Bernasconi, Neda Bernasconi, Alexander Weil, Raluca Pana, et al. Differential reorganization of episodic and semantic memory systems in epilepsy-related mesiotemporal pathology. Brain, 147(11):3918–3932, 2024.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. Findings of the association for computational linguistics: ACL 2024, pp. 13590–13618, 2024.

Chengkuan Chen, Luca L Weishaupt, Drew FK Williamson, Richard J Chen, Tong Ding, Bowen Chen, Anurag Vaidya, Long Phi Le, Guillaume Jaume, Ming Y Lu, et al. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. arXiv preprint arXiv:2506.20964, 2025a.

Haizhu Chen, Ruichong Lin, and Yu Yunfang. Metapath chat: multimodal generative artificial intelligence chatbot for clinical pathology. MedComm, 5(10):e769, 2024.

Jingyun Chen, Linghan Cai, Zhikang Wang, Yi Huang, Songhan Jiang, Shenjin Huang, Hongpeng Wang, and Yongbing Zhang. Pathagent: Toward interpretable analysis of whole-slide pathology images via large language model-based agentic reasoning. arXiv preprint arXiv:2511.17052, 2025b.

Pingyi Chen, Chenglu Zhu, Sunyi Zheng, Honglin Li, and Lin Yang. Wsi-vqa: Interpreting whole slide images by generative visual question answering. In European Conference on Computer Vision, pp. 401–417. Springer, 2025c.

Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 5134–5143, 2025d.

Didem Cifci, Gregory P Veldhuizen, Sebastian Foersch, and Jakob Nikolas Kather. Ai in computational pathology of cancer: improving diagnostic workflows and clinical outcomes? Annual Review of Cancer Biology, 7(1):57–71, 2023.

Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. Nature medicine, pp. 1–13, 2025.

Cody V Dong, Qihong Lu, Kenneth A Norman, and Sebastian Michelmann. Towards large language models with human-like episodic memory. Trends in Cognitive Sciences, 2025.

Anna Fink, Alexander Rau, Marco Reisert, Fabian Bamberg, and Maximilian F Russe. Retrieval-augmented generation with large language models in radiology: from theory to practice. Radiology: Artificial Intelligence, 7(4):e240790, 2025.

Elena Fountzilas, Tillman Pearce, Mehmet A Baysal, Abhijit Chakraborty, and Apostolia M Tsimberidou. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. NPJ Digital Medicine, 8(1):75, 2025.

Mariano I Gabitto, Kyle J Travaglini, Victoria M Rachleff, Eitan S Kaplan, Brian Long, Jeanelle Ariza, Yi Ding, Joseph T Mahoney, Nick Dee, Jeff Goldy, et al. Integrated multimodal cell atlas of alzheimer's disease. Nature neuroscience, 27(12):2366–2383, 2024.

Advait Gosai, Arun Kavishwar, Stephanie L McNamara, Soujanya Samineni, Renato Umeton, Alexander Chowdhury, and William Lotter. Beyond diagnosis: Evaluating multimodal llms for pathology localization in chest radiographs. arXiv preprint arXiv:2509.18015, 2025.

Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Liansheng Wang, and Hao Chen. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 189–199. Springer, 2024.

Ameer Hamza, Yong Hyun Ahn, Sungyoung Lee, Seong Tae Kim, et al. Llava needs more knowledge: Retrieval augmented natural language generation with knowledge graph for explaining thoracic pathologies. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 3311–3319, 2025.

Quanhao He, Hao Tan, Bangxin Xiao, Yiwen Tan, Xiang Peng, Canjie Peng, Xiaofeng Yue, Linshan Jiang, Youde Cao, Fa Jin Lv, et al. Prior knowledge-guided multimodal deep learning system for biomarker exploration and prognosis prediction of urothelial carcinoma. npj Digital Medicine, 2025.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Yan Hu, Wenjie Li, and Jiang Liu. Memory-augmented multimodal llms for surgical vqa via self-contained inquiry. arXiv preprint arXiv:2411.10937, 2024.

Yuki Hou, Haruki Tamoto, Qinghua Zhao, and Homei Miyashita. Synapticrag: Enhancing temporal memory retrieval in large language models through synaptic mechanisms. In Findings of the Association for Computational Linguistics: ACL 2025, pp. 20422–20436, 2025.

Qingqiao Hu, Weimin Lyu, Meilong Xu, Kehan Qi, Xiaoling Hu, Saumya Gupta, Jiawei Zhou, and Chao Chen. Loc-path: Learning to compress for pathology multimodal large language models. arXiv preprint arXiv:2512.05391, 2025.

Weiquan Huang, Zixuan Wang, Hehai Lin, Sudong Wang, Bo Xu, Qian Li, Beier Zhu, Linyi Yang, and Chengwei Qin. Ama: Adaptive memory via multi-agent collaboration. arXiv preprint arXiv:2601.20352, 2026.

Yanyan Huang, Weiqin Zhao, Zhengyu Zhang, Yihang Chen, Yu Fu, Feng Wu, Yuming Jiang, Li Liang, Shujun Wang, and Lequan Yu. Knowledge-guided adaptation of pathology foundation models effectively improves cross-domain generalization and demographic fairness. Nature Communications, 2025.

Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm. Authorea Preprints, 2024.

Mohamed Sobhi Jabal, Pranav Warman, Jikai Zhang, Kartikeye Gupta, Ayush Jain, Maciej Mazurowski, Walter Wiggins, Kirti Magudia, and Evan Calabrese. Open-weight language models and retrieval-augmented generation for automated structured data extraction from diagnostic reports: assessment of approaches and parameters. Radiology: Artificial Intelligence, 7(3):e240551, 2025.

Hanqi Jiang, Junhao Chen, Yi Pan, Ling Chen, Weihang You, Yifan Zhou, Ruidong Zhang, Andrea Sikora, Lin Zhao, Yohannes Abate, et al. Synapse: Empowering llm agents with episodic-semantic memory via spreading activation. arXiv preprint arXiv:2601.02744, 2026a.

Rui Jiang, Xiaoxu Yin, Pengshuai Yang, Lingchao Cheng, Juan Hu, Jiao Yang, Ying Wang, Xiaodan Fu, Li Shang, Liling Li, et al. A transformer-based weakly supervised computational pathology method for clinical-grade diagnosis and molecular marker discovery of gliomas. Nature Machine Intelligence, 6(8):876–891, 2024.

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, 2026b.

Maria Eleni Karakatsani, Robin Ji, Maria F Murillo, Tara Kugelman, Nancy Kwon, Yeh-Hsing Lao, Keyu Liu, Antonios N Pouliopoulos, Lawrence S Honig, Karen E Duff, et al. Focused ultrasound mitigates pathology and improves spatial memory in alzheimer's mice and patients. Theranostics, 13(12):4102, 2023.

Sangwook Kim, Soonyoung Lee, and Jongseong Jang. Chatexaonepath: An expert-level multimodal large language model for histopathology using whole slide images. arXiv preprint arXiv:2504.13023, 2025.

Dong Li, Guihong Wan, Xintao Wu, Xinyu Wu, Xiaohui Chen, Yi He, Zhong Chen, Peter K Sorger, and Chen Zhao. Multi-modal foundation models for computational pathology: a survey. Transactions on machine learning research, 2025:5715, 2025.

Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, et al. Wsi-llava: A multimodal large language model for whole slide image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22718–22727, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916, 2023.

Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, et al. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 9087–9097, 2025a.

Yu Liu, Xixun Lin, Yanmin Shang, Yangxi Li, Shi Wang, and Yanan Cao. Pathmind: A retrieve-prioritize-reason framework for knowledge graph reasoning with large language models. arXiv preprint arXiv:2511.14256, 2025b.

Yunsong Liu, Zunamys I Carrero, Xiaofeng Jiang, Dyke Ferber, Georg Wölflein, Li Zhang, Sanddhya Jayabalan, Tim Lenz, Zhouguang Hui, and Jakob Nikolas Kather. Benchmarking large language model-based agent systems for clinical decision tasks. npj Digital Medicine, 2026.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. Nature, 634(8033):466–473, 2024.

Jiabo Ma, Zhengrui Guo, Fengtao Zhou, Yihui Wang, Yingxue Xu, Jinbang Li, Fang Yan, Yu Cai, Zhengjie Zhu, Cheng Jin, et al. A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. Nature Biomedical Engineering, pp. 1–20, 2025.

Shaffi Manchanda, Lorena Galan-Acosta, Axel Abelein, Simone Tambaro, Gefei Chen, Per Nilsson, and Jan Johansson. Intravenous treatment with a molecular chaperone designed against β-amyloid toxicity improves alzheimer's disease pathology in mouse models. Molecular Therapy, 31(2): 487–502, 2023.

Paras S Minhas, Jeffrey R Jones, Amira Latif-Hernandez, Yuki Sugiura, Aarooran S Durairaj, Qian Wang, Siddhita D Mhatre, Takeshi Uenaka, Joshua Crapser, Travis Conley, et al. Restoring hippocampal glucose metabolism rescues cognition across alzheimer's disease pathologies. Science, 385(6711):eabm6131, 2024.

Mohamed Omar, Varun Ullanat, Massimo Loda, Luigi Marchionni, and Renato Umeton. Chatgpt for digital pathology research. The Lancet Digital Health, 6(8):e595–e600, 2024.

Raymond W Redline, Drucilla J Roberts, Mana M Parast, Linda M Ernst, Terry K Morgan, Michael F Greene, Cynthia Gyamfi-Bannerman, Judette M Louis, Emin Maltepe, Karen K Mestan, et al. Placental pathology is necessary to understand common pregnancy complications and achieve an improved taxonomy of obstetrical disease. American journal of obstetrics and gynecology, 228 (2):187–202, 2023.

Roberto Romero, Eunjung Jung, Tinnakorn Chaiworapongsa, Offer Erez, Dereje W Gudicha, Yeon Mee Kim, Jung-Sun Kim, Bomi Kim, Juan Pedro Kusanovic, Francesca Gotsch, et al. Toward a new taxonomy of obstetrical disease: improved performance of maternal blood biomarkers for the great obstetrical syndromes when classified according to placental pathology. American journal of obstetrics and gynecology, 227(4):615–e1, 2022.

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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13183–13192, 2024.

Abubakr Shafique, Ricardo Gonzalez, Liron Pantanowitz, Puay Hoon Tan, Alberto Machado, Ian A Cree, and Hamid R Tizhoosh. A preliminary investigation into search and matching for tumor discrimination in world health organization breast taxonomy using deep networks. Modern Pathology, 37(2):100381, 2024.

George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. arXiv preprint arXiv:2405.10254, 2024.

Jintao Sheng, Alexandra N Trelle, America Romero, Jennifer Park, Tammy T Tran, Sharon J Sha, Katrin I Andreasson, Edward N Wilson, Elizabeth C Mormino, and Anthony D Wagner. Top-down attention and alzheimer's pathology affect cortical selectivity during learning, influencing episodic memory in older adults. Science advances, 11(24):eads4206, 2025.

Jiangbo Shi, Chen Li, Yefeng Zheng, and Huazhu Fu. Kpvg: Knowledge-prompted vision-genomics model for cancer survival prediction in whole slide images. Information Fusion, pp. 103660, 2025.

JF Smart and M Roux. A model for medical knowledge representation application to the analysis of descriptive pathology reports. Methods of information in medicine, 34(04):352–360, 1995.

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, Yixuan Si, Chenglu Zhu, Xuan Gong, Kai Zhang, Pingyi Chen, Ye Zhang, Zhongyi Shui, Tao Lin, and Lin Yang. Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 10360–10371, 2025.

Ehsan Ullah, Mirza Mansoor Baig, Asim Waqas, Ghulam Rasool, Rajendra Singh, Ashwinikumar Shandilya, Hamid GholamHossieni, and Anil V Parwani. Multimodal generative ai for anatomic pathology—a review of current applications to envisage the future direction. Advances in anatomic pathology, pp. 10–1097, 2025.

Gregory Verghese, Jochen K Lennerz, Danny Ruta, Wen Ng, Selvam Thavaraj, Kalliopi P Siziopikou, Threnesan Naidoo, Swapnil Rane, Roberto Salgado, Sarah E Pinder, et al. Computational pathology in cancer diagnosis, prognosis, and prediction–present day and prospects. The Journal of pathology, 260(5):551–563, 2023.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. Nature, pp. 1–8, 2024.

Luodan Yang, Chongyun Wu, Yong Li, Yan Dong, Celeste Yin-Chieh Wu, Reggie Hui-Chao Lee, Darrell W Brann, Hung Wen Lin, and Quanguang Zhang. Long-term exercise pre-training attenuates alzheimer's disease–related pathology in a transgenic rat model of alzheimer's disease. Geroscience, 44(3):1457–1477, 2022.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. National Science Review, 11(12):nwae403, 2024.

Ryo Yoshida, Shinnosuke Isono, Kohei Kajikawa, Taiga Someya, Yushi Sugimoto, and Yohei Oseki. If attention serves as a cognitive model of human memory retrieval, what is the plausible memory representation? In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9795–9812, 2025.

Wenchuan Zhang, Jingru Guo, Hengzhe Zhang, Penghao Zhang, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. Patho-agenticrag: towards multimodal agentic retrieval-augmented generation for pathology vlms via reinforcement learning. arXiv preprint arXiv:2508.02258, 2025.

Weiqin Zhao, Ziyu Guo, Yinshuang Fan, Yuming Jiang, Maximus CF Yeung, and Lequan Yu. Aligning knowledge concepts to whole slide images for precise histopathology image analysis. npj Digital Medicine, 7(1):383, 2024.

A Long-term Memory Construction

To construct the pathology-oriented knowledge graph, we design an evidence-driven deep-search pipeline that automatically mines disease-relevant knowledge from PubMed abstracts (As shown in Algorithm 1). Given a predefined disease query, the pipeline first retrieves PubMed identifiers and fetches their corresponding XML records through the NCBI Entrez interface. Abstracts without valid textual content are discarded, and duplicate abstracts are removed by normalized hashing. Each remaining abstract is processed by a large language model using a constrained pathology schema, including disease entities, anatomical sites, histology, morphologic features, immunohistochemical markers, molecular alterations, serum markers, and diagnostic clues. The extraction module is instructed to output only explicitly supported information with evidence spans and confidence scores. To improve reliability, each extracted relation is further normalized, confidence-calibrated, and filtered by a minimum confidence threshold before being inserted into the knowledge graph. The final graph is saved both as plain triples and as rich evidence-aware edges, while disease-centric node memories and feature-to-disease indices are exported for subsequent retrieval and model guidance.

B More Implementation Details

B.1 Dataset & Processing

Experiments are mainly conducted on **WSI-Bench**, with additional zero-shot studies on independent datasets to evaluate generalization ability. **In-domain benchmark.** We use **WSI-Bench** as the core benchmark for both model development and quantitative evaluation. Derived from TCGA, it focuses on pathology-specific visual reasoning over whole-slide images (WSIs). The benchmark consists of 9,850 WSIs spanning 30 tumor types and includes 179,569 associated VQA instances. The annotations are designed to evaluate four key aspects of pathological intelligence: (1) morphology understanding, (2) diagnostic assessment, (3) therapy-related reasoning, and (4) report-level text generation. Both generative and multiple-choice answer formats are included. We strictly follow the official data split without modification. Specifically, 9,642 WSIs with 175,450 QA pairs are used for training, while 208 WSIs with 4,119 QA pairs are reserved for evaluation. All reported metrics follow the original benchmark protocol, including WSI-P, WSI-R, classification accuracy, BLEU, and ROUGE. **Out-of-domain benchmarks.** To evaluate generalization to unseen data distributions, we conduct zero-shot experiments on three independent datasets: *WSI-VQA* (Chen et al., 2025c), *SlideBench-VQA (BCNB)* (Chen et al., 2025d), and *CPTAC-NSCLC*. Following prior work, we apply data-cleaning procedures to remove samples that are not clinically well-posed or cannot be answered based on WSI content. This includes questions involving prognosis prediction, unavailable biomarker information, or three-dimensional lesion size estimation. Importantly, these external datasets are used strictly for evaluation purposes, and no additional fine-tuning or adaptation is performed.

B.2 Implementation Details

WSI Representation Pipeline. Given the gigapixel nature of whole-slide images, we adopt a hierarchical processing strategy to balance efficiency and contextual awareness. Each WSI is first divided into non-overlapping tiles of size 256×256 at a predefined magnification level. These tiles are independently encoded using a visual backbone to produce patch-level feature embeddings, following WSI-LLaVA (Liang et al., 2025). To capture global structural information, the patch embeddings are aggregated and passed through a LongNet-based transformer. This module is designed to model long-range dependencies across the entire slide, enabling effective representation of spatial relationships that extend beyond local regions. The final visual representation is projected into the language embedding space via a trainable linear layer, which serves as the interface between vision and language modalities. **Language Modeling and Multimodal Fusion.** We leverage a pretrained large language model (LLM) as the textual backbone for downstream reasoning and generation tasks. The projected visual features are aligned with token embeddings and injected into the LLM through a multimodal projector. This design allows the model to condition text generation on slide-level visual context. Depending on the training stage, different subsets of parameters are frozen to stabilize optimization and reduce computational cost. **Multi-Stage Optimization.** Training is organized into three sequential stages: ❶ **Cross-Modal Pre-alignment.** We first perform contrastive learning between slide-level visual embeddings and corresponding pathology reports to establish a shared

Algorithm 1: Pseudocode of the pathology knowledge graph construction pipeline.

Input: Disease-specific PubMed query Q; maximum number of articles N; confidence threshold τ; LLM extraction schema $\mathcal{S}$
Output: Pathology knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; triples $\mathcal{T}$; disease memory $\mathcal{M}_d$; feature index $\mathcal{M}_f$

Initialize empty pathology KG $\mathcal{G}$ and triple set $\mathcal{T}$
Initialize seen abstract hash set $\mathcal{H}$
/* PubMed deep search */
$\mathcal{P} \leftarrow \mathrm{PubMedSearch}(Q, N)$
$\mathcal{A} \leftarrow \mathrm{PubMedFetch}(\mathcal{P})$
foreach article $a \in \mathcal{A}$ **do**
  Extract PMID m and abstract text x from a
  if x is empty **then**
    continue
  end
  /* Abstract-level deduplication */
  $h \leftarrow \mathrm{Hash}(\mathrm{Normalize}(x))$
  if $h \in \mathcal{H}$ **then**
    continue
  end
  $\mathcal{H} \leftarrow \mathcal{H} \cup \{h\}$
  /* Schema-constrained LLM extraction */
  $r \leftarrow \mathrm{LLMExtract}(x, \mathcal{S})$
  Save raw extraction (m, r) for audit and retry
  if r is invalid **then**
    continue
  end
  Let d be the normalized disease name in r
  if d is empty **then**
    continue
  end
  /* Relation normalization and graph insertion */
  foreach candidate item (s, p, o, e, c) extracted from r **do**
    $p' \leftarrow \mathrm{SlugRelation}(p)$
    $c' \leftarrow \mathrm{CalibrateConfidence}(c, e)$
    $e' \leftarrow \mathrm{ClipEvidence}(e)$
    if $c' \geq \tau$ and s, o, p' are valid **then**
      Add triple (s, p', o) into $\mathcal{T}$
      Add rich edge (s, p', o, m, c', e') into $\mathcal{E}$
      Update node set $\mathcal{V} \leftarrow \mathcal{V} \cup \{s, o\}$
    end
  end
end
/* Memory export */
Save $\mathcal{T}$ as `triples.tsv`
Save evidence-aware edges $\mathcal{E}$ as `edges.jsonl`
$\mathcal{M}_d \leftarrow \mathrm{ExportDiseaseNodes}(\mathcal{T})$
$\mathcal{M}_f \leftarrow \mathrm{ExportFeatureIndex}(\mathcal{T})$
return $\mathcal{G}, \mathcal{T}, \mathcal{M}_d, \mathcal{M}_f$

embedding space. ❷ **Projection Adaptation.** Next, we connect the pretrained visual encoder with the LLM through a projection layer. During this stage, both the visual encoder and LLM remain frozen, and only the projection parameters are updated. ❸ **Instruction Fine-tuning.** Finally, we conduct task-specific training on WSI-Bench. The projection layer and LLM are jointly optimized, while the visual encoder is kept fixed to preserve learned visual representations. **Evaluation Metrics.** We evaluate model performance from both linguistic and clinical perspectives. For open-ended generation tasks, BLEU, ROUGE-L, and METEOR are used to assess lexical overlap and semantic similarity with reference reports. To better reflect pathology-specific correctness, we additionally employ WSI-P and WSI-R, which measure the accuracy of predicted findings and their consistency with visual evidence, respectively. **Training Configuration.** All experiments are conducted using DeepSpeed on 4 NVIDIA A100 GPUs. Optimization is performed with AdamW, using an initial learning rate of 2×10^{-4} and a cosine decay schedule with a warmup ratio of 0.03. The model is trained for 3 epochs with a per-device batch size of 32. For parameter-efficient adaptation, we adopt LoRA with rank $r = 128$ and scaling factor $\alpha = 256$. The multimodal projection layer is trained with a smaller learning rate of 2×10^{-5} to ensure stable alignment between modalities.

C Evaluation Prompts

To enable fine-grained and more reliable evaluation of model-generated pathology outputs, we design a multi-stage prompting pipeline for both **WSI-R (WSI-R)** and **WSI-P (WSI-P)**. Our evaluation operates at the *claim level*, where free-form text is decomposed into atomic factual units and subsequently assessed using large language models (LLMs). ❶ **Claim Extraction.** We first transform long-form pathology text into structured atomic claims. **WSI-R.** For WSI-R, claims are extracted from model-generated diagnostic text. The prompt instructs the LLM to decompose the text into granular, non-overlapping claims while preserving semantic coherence. Closely related information is kept within the same claim to avoid breaking contextual dependencies. The model is required to avoid omission or duplication and to output results strictly as a list of claims. **WSI-P.** For WSI-P, claim extraction is conditioned on question-answer pairs. The prompt additionally filters out information that is not directly relevant to the question and retains only claims that explicitly address the query. This ensures that subsequent evaluation focuses on answer-specific correctness rather than general content. ❷ **Claim-level Evaluation.** After claim extraction, we evaluate each claim using task-specific prompts. **WSI-R Evaluation.** In WSI-R, each model-generated claim is evaluated against the ground truth answer. The LLM acts as an impartial judge and assigns a graded relevance score along with a natural language explanation. The scoring rubric is defined as follows: **1.0**: Fully supported by the ground truth **0.7**: Mostly supported with minor omissions **0.3**: Partially supported with significant gaps **0.0**: Not supported Claims that are not addressed in the ground truth are ignored to avoid penalizing missing coverage. **WSI-P Evaluation.** In WSI-P, evaluation focuses on factual correctness. Each ground truth claim is compared against the model response, and the LLM assigns a correctness score with explanation. The scoring rubric is: **1.0**: Completely correct **0.7**: Mostly correct **0.3**: Contains core factual errors **0.0**: Completely incorrect This formulation emphasizes clinically meaningful correctness, including subtle diagnostic inaccuracies. ❸ **Output Format.** Both evaluation prompts require the LLM to output results as a list of structured entries, where each entry contains: the claim, an explanation, and a numerical score. This standardized format facilitates reliable parsing and downstream aggregation. ❹ **Score Aggregation** Given claim-level scores, we compute: **Per-sample score**: average across all claims in a sample, **Type-level score**: average within each metadata category, **Overall score**: average across all samples. Invalid or missing claims are excluded from aggregation to ensure robustness. ❺ **Design Rationale** Our prompting design is guided by three principles: **Decomposition**: Breaking complex pathology text into atomic claims reduces ambiguity. **Separation of Concerns**: WSI-R evaluates relevance (coverage), while WSI-P evaluates correctness. **Graded Scoring**: Continuous scores provide more nuanced evaluation than binary metrics.

D Compute Reporting

All experiments were conducted using `DeepSpeed` on 4 NVIDIA A100 GPUs. The model was optimized with AdamW using an initial learning rate of 2×10^{-4}, together with a cosine decay learning rate schedule and a warmup ratio of 0.03. Training was performed for 3 epochs with a

per-device batch size of 32. For parameter-efficient fine-tuning, we adopted LoRA with rank $r = 128$ and scaling factor $\alpha = 256$. To stabilize multimodal alignment, the multimodal projection layer was optimized with a smaller learning rate of 2×10^{-5}. The overall compute budget was moderate due to the use of parameter-efficient adaptation. Compared with full-parameter fine-tuning, LoRA significantly reduced the number of trainable parameters and GPU memory consumption, enabling all experiments to be completed on a 4-GPU setup. In addition, mixed-precision training and distributed optimization provided by `DeepSpeed` were used to further improve computational efficiency and reduce training overhead.

E More Qualitative Case Study

To further illustrate the behavior of `PathMem`, we present a set of qualitative case studies as shown in Figure 5, Figure 6, Figure 7, Figure 8, Figure 9. These examples collectively provide an intuitive comparison across different methods and evaluation perspectives, highlighting variations in factual correctness, information completeness, and semantic alignment with the reference reports. Overall, the visualized results demonstrate that `PathMem` produces more accurate and comprehensive descriptions, while maintaining better consistency with domain knowledge and reducing template-like or spurious content compared to baseline approaches.

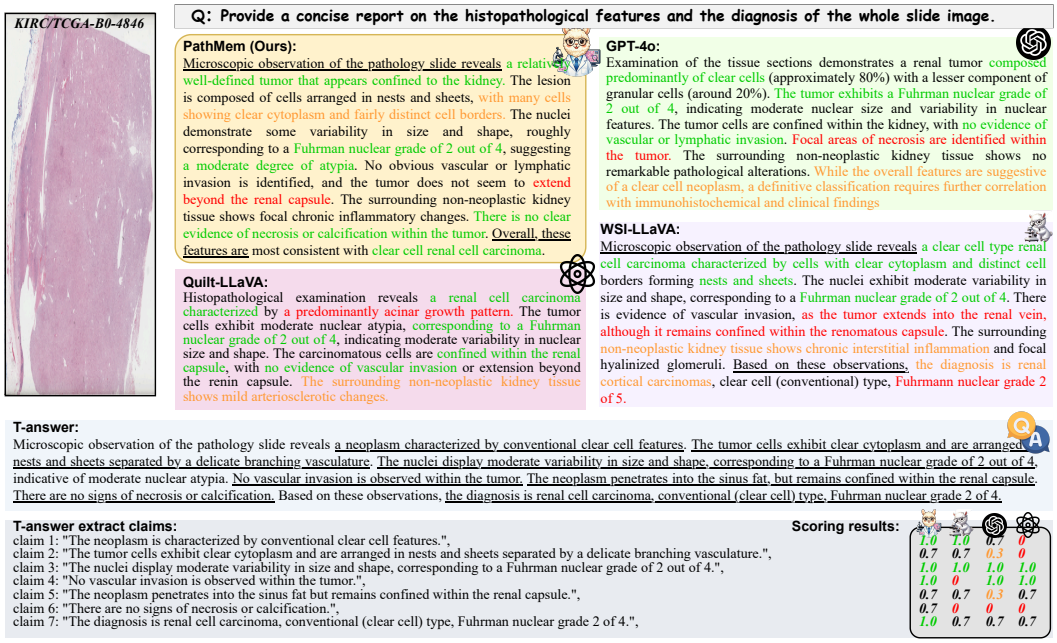

Figure 5: Qualitative comparison of generated reports by our method and three baseline approaches on the report generation task. (Red highlights denote incorrect content, while green highlights denote correct content, and orange highlights denote missing content from T-answer.)

Model	Model Responses	WSI Metrics		NLU Metrics
		WSI-P	WSI-R	
Ground Truth	Microscopic observation of the pathology slide reveals poorly differentiated clusters of adenocarcinoma cells exhibiting significant architectural disorganization and extensive necrosis. There is an evident absence of glandular structure formation, indicative of poor differentiation. The slide also shows notable angiolymphatic invasion and evidence of visceral pleural invasion, as the tumor has penetrated the pleural surface. Additionally, the presence of multiple satellite nodules further exemplifies the invasive nature of this carcinoma. Background lung tissue displays changes consistent with emphysema. Based on these observations, the diagnosis is invasive poorly differentiated adenocarcinoma.	N/A	N/A	BLEU-1: — BLEU-2: — BLEU-3: — BLEU-4: — ROUGE-L: — METEOR: —
Hist-Gen	Histological evaluation demonstrates an aggressive epithelial malignancy of gland-forming origin with limited differentiation. The neoplasm is largely composed of sheets of tumor cells arranged in a compact pattern, while a minor proportion displays focal glandular architecture. Many cells contain abundant intracellular mucin that compresses the nuclei toward the cell membrane, producing a signet-ring–like appearance. The supporting stroma is dense with areas of fibrosis and hyaline change. No vascular or lymphatic permeation is identified. Adjacent pulmonary parenchyma exhibits features of chronic airspace enlargement, and a small neuroendocrine tumorlet is noted away from the primary lesion. Overall findings are consistent with a high-grade pulmonary adenocarcinoma.	0.257	0.230	BLEU-1: 0.274 BLEU-2: 0.100 BLEU-3: 0.045 BLEU-4: 0.025 ROUGE-L: 0.182 METEOR: 0.204
WSI-LLaVA	The examined lung specimen reveals a malignant glandular tumor with marked loss of structural organization. Most of the lesion consists of densely packed tumor cells, whereas only a small fraction forms rudimentary acinar structures. Numerous cells demonstrate intracellular mucous accumulation with peripheral displacement of nuclei. The tumor microenvironment shows prominent stromal sclerosis. Invasion into vascular or lymphatic channels is not observed. Background lung tissue is notable for airspace destruction compatible with emphysema, and a minute carcinoid proliferation is present at a distance from the main mass. These morphological features support the interpretation of a poorly differentiated carcinoma of adenocarcinomatous type arising in the lung.	0.311	0.294	BLEU-1: 0.319 BLEU-2: 0.141 BLEU-3: 0.071 BLEU-4: 0.043 ROUGE-L: 0.250 METEOR: 0.258
PathMem	Microscopic examination of the pathology slide reveals poorly differentiated clusters of adenocarcinoma cells with a predominant solid growth pattern and marked architectural disorganization. There is minimal glandular structure formation, consistent with poor differentiation, and focal areas show necrotic change. The tumor demonstrates evidence of angiolymphatic invasion, and focal visceral pleural involvement is identified, indicating penetration toward the pleural surface. Scattered small nodular extensions are present within the surrounding parenchyma, supporting the invasive behavior of the carcinoma. Background lung tissue shows changes consistent with emphysema. Based on these findings, the diagnosis is invasive poorly differentiated adenocarcinoma.	0.596	0.592	BLEU-1: 0.636 BLEU-2: 0.519 BLEU-3: 0.426 BLEU-4: 0.361 ROUGE-L: 0.583 METEOR: 0.678

Figure 6: Comparison of WSI-based and NLU-based evaluations. Green indicates agreement with the reference, red denotes deviations, orange marks missing ground-truth content, underlined text reflects template-style language, and blue highlights knowledge graph–grounded concepts.

Model	Model Responses	WSI Metrics		NLU Metrics
		WSI-P	WSI-R	
Ground Truth	Microscopic observation of the pathology slide reveals an epithelioid malignant tumor comprising approximately 80% of the examined tissue, with sparse connective tissue presence. The tumor cells are characterized by round to oval nuclei with granular chromatin and eosinophilic nucleoli, accompanied by a high nucleus-to-cytoplasm ratio. The cells display a mitotic count of 10 mitoses per square millimeter, indicative of a high cell division rate. The cytoplasm of the tumor cells is of moderate size, and the tumor exhibits a sheet-like growth pattern. Approximately 10% of the tumor tissue is necrotic. Lymphocytic infiltrates make up about 3% of the tissue cells. Based on these observations, the diagnosis is consistent with a metastasis of a melanoma.	N/A	N/A	BLEU-1: — BLEU-2: — BLEU-3: — BLEU-4: — ROUGE-L: — METEOR: —
Hist-Gen	The slide shows a malignant neoplasm composed mainly of poorly differentiated epithelioid cells with minimal intervening stroma. The cells have enlarged hyperchromatic nuclei and limited cytoplasm, and focal pigment is present within the lesion. Growth is predominantly diffuse, with visible proliferative activity and scattered degenerative or inflammatory changes in the background. Overall, the appearance favors a metastatic high-grade melanocytic tumor, consistent with melanoma.	0.437	0.522	BLEU-1: 0.205 BLEU-2: 0.154 BLEU-3: 0.083 BLEU-4: 0.042 ROUGE-L: 0.229 METEOR: 0.177
WSI-LLaVA	Microscopic observation of the pathology slide reveals an epithelioid malignant tumor comprising approximately 80% of the tissue, with sparse connective tissue presence. The tumor cells exhibit round to oval nuclei with granular chromatin and have a high nucleocytoplasmic ratio. The cytoplasm is moderately pigmented, and there is a sheet-like growth pattern observed. A mitotic count reveals 8 mitoses per square millimeter, indicating a high rate of cell division. Presence of apoptotic cells and lymphoc	0.500	0.589	BLEU-1: 0.461 BLEU-2: 0.396 BLEU-3: 0.348 BLEU-4: 0.311 ROUGE-L: 0.570 METEOR: 0.432
PathMem	Microscopic observation of the pathology slide reveals an epithelioid malignant tumor comprising approximately 95% of the tissue, with sparse connective tissue presence. The tumor cells exhibit round to oval nuclei with granular chromatin and have a high nucleocyto/cytoplasm ratio. The cytoplasms of these cells are moderately pigmented, and there is a sheet-like growth pattern observed. A mitotic count reveals 8 mitoses per square millimeter, indicating a high rate of cell division. Presence of apoptotic cells and lymphocytic infiltrates constitute about 3% of cells in the tissues, along with melanophages. Based on these observations, the diagnosis is a metastasis of a melanoma.	0.700	0.633	BLEU-1: 0.700 BLEU-2: 0.585 BLEU-3: 0.501 BLEU-4: 0.438 ROUGE-L: 0.719 METEOR: 0.698
KG Activation	idx: 7318 metastatic melanoma → HAS_DIAGNOSTIC_CLUE → TRB V-CDR3-J intrinsic disorder diversity（novel prognostic/diagnostic biomarker reflecting metastatic progression） idx: 11643 uveal melanoma → HAS_DIAGNOSTIC_CLUE → presence of metastatic lesions (direct evidence of metastatic melanoma)	Immune-related **biomarkers** for metastatic melanoma The presence of metastatic lesions is itself important **diagnostic evidence** for melanoma.		

Figure 7: Comparison of WSI-based and NLU-based evaluations. Green indicates agreement with the reference, red denotes deviations, orange marks missing ground-truth content, underlined text reflects template-style language, and blue highlights knowledge graph–grounded concepts.

Model	Model Responses	WSI Metrics		NLU Metrics
		WSI-P	WSI-R	
Ground Truth	The whole slide image reveals a predominantly epithelioid malignant tumor occupying approximately 80% of the examined tissue. The tumor exhibits a sheet-like growth pattern with sparse connective tissue. The cells display round to oval nuclei with granular chromatin and eosinophilic nucleoli. There is a notable high nucleus-to-cytoplasm ratio, and the cytoplasm is of moderate size. Additionally, around 10% of the tissue is necrotic, and lymphocytic infiltrates are present, comprising about 3% of the cells.	N/A	N/A	BLEU-1: — BLEU-2: — BLEU-3: — BLEU-4: — ROUGE-L: — METEOR: —
Hist-Gen	The examined section is largely replaced by an aggressive tumor composed of epithelial-like cells, with only a limited amount of supporting tissue remaining. The neoplastic cells vary slightly in shape but tend to appear rounded, and their nuclei occupy much of the cell volume. Nuclear features are somewhat coarse, and cell division can be observed at multiple points. The tumor grows in broad clusters rather than forming distinct structures. In addition, small numbers of inflammatory cells are interspersed within the lesion.	0.500	0.518	BLEU-1: 0.356 BLEU-2: 0.155 BLEU-3: 0.065 BLEU-4: 0.035 ROUGE-L: 0.191 METEOR: 0.225
WSI-LLaVA	The whole slide image shows a predominance of epithelioid malignant tumor cells occupying roughly 90% of the tissue. These cells form a dense sheet-like pattern with minimal intervening connective tissue, indicating a high cellularity. The nuclei are round to oval with granular chromatin and a high nucleus-to-cytoplasm ratio, suggesting aggressive cellular behavior. The mitotic rate is observed at 3 mitoses per square millimeter. A small amount of the tumor tissue (5%) exhibits lymphocytic infiltrates, while another 5% is necrotic.	0.571	0.550	BLEU-1: 0.630 BLEU-2: 0.408 BLEU-3: 0.264 BLEU-4: 0.170 ROUGE-L: 0.532 METEOR: 0.549
PathMem	The whole slide image reveals an epithelioid malignant tumor occupying approximately 90 percent of the tissue, with very little connective tissue observed. The tumor cells are predominantly round to oval shaped, exhibiting granular chromatin and a high nucleo/cytoplasmic ratio. There is a significant presence of mitotic figures, with 8 mitoses per square millimeter. The cytoplasm of these tumor cell is bipolar and the tumor displays a sheet-like growth pattern. Additionally, lymphocytic infiltrates are scattered throughout the tissues.	0.629	0.713	BLEU-1: 0.674 BLEU-2: 0.480 BLEU-3: 0.353 BLEU-4: 0.237 ROUGE-L: 0.588 METEOR: 0.554
KG Activation	idx: 11851 uveal melanoma → HAS_HISTOLOGIC_TYPE → epithelioid idx: 6766 perivascular epithelioid cell tumor → HAS_HISTOLOGIC_TYPE → epithelioid neoplasm idx: 6769 perivascular epithelioid cell tumor → HAS_CELLULAR_FEATURE → dispersed and loosely cohesive epithelioid cells idx: 6770 perivascular epithelioid cell tumor → HAS_CELLULAR_FEATURE → enlarged round to oval nuclei idx: 6771 perivascular epithelioid cell tumor → HAS_CELLULAR_FEATURE → fine chromatin			"**epithelioid malignant tumor**" can support the direction of melanoma-related diagnosis. supports the identification of the morphological classification of the broad category of **"epithelioid tumor"** Structural analysis supporting **the distribution of tumor cells in sheets** Directly corresponds to "**round to oval shaped**", enhancing the matching of cell morphology features. The corresponding term "**granular chromatin**" supports the determination of nuclear chromatin morphology.

Figure 8: Comparison of WSI-based and NLU-based evaluations. Green indicates agreement with the reference, red denotes deviations, orange marks missing ground-truth content, underlined text reflects template-style language, and blue highlights knowledge graph–grounded concepts.

Model	Model Responses	WSI Metrics		NLU Metrics
		WSI-P	WSI-R	
Ground Truth	Microscopic observation of the pathology slide reveals exuberant granulation tissue with characteristic fibroblastic proliferation and myxoid changes, typical for a high-grade myxofibrosarcoma. No residual sarcoma tissue is identified in the examined section, suggesting a primarily reactive tissue response. The tissue is notably devoid of vascular invasion, calcification, necrosis, or significant atypia or pleomorphism within the granulation tissue. Based on these observations, the diagnosis is myxofibrosarcoma	N/A	N/A	BLEU-1: — BLEU-2: — BLEU-3: — BLEU-4: — ROUGE-L: — METEOR: —
Hist-Gen	The examined section predominantly shows a reactive stromal process rather than overt malignant tissue. There is prominent loose extracellular matrix with scattered spindle-shaped cells, consistent with a reparative or post-surgical change. Features typically associated with aggressive sarcoma, such as marked cellular atypia or destructive growth, are not clearly evident in this field. The presence of foreign material suggests prior intervention, which may account for the observed tissue response. Overall, the findings are in keeping with a treated or previously diagnosed soft tissue sarcoma, most consistent with myxofibrosarcoma in context.	0.562	0.689	BLEU-1: 0.320 BLEU-2: 0.150 BLEU-3: 0.061 BLEU-4: 0.032 ROUGE-L: 0.272 METEOR: 0.292
WSI-LLaVA	Microscopic observation of the pathology slide reveals exuberant granulation tissue with characteristic fibroblastic proliferation and myxoid changes, typical for a high-grade myxofibrosarcoma. No residual sarcoma tissue is identified in the examined section, suggesting a primarily reactive tissue response. The tissue architecture is notably altered by the presence of foreign material, as indicated by the negative tumor section. Based on these observations, the diagnosis is myxofibrosarcoma, histologic grade 3 according to the FNCLCC grading system, reflecting poor tumor differentiation, significant mitotic activity, and the presenceof necrosis.	0.821	0.625	BLEU-1: 0.650 BLEU-2: 0.593 BLEU-3: 0.566 BLEU-4: 0.549 ROUGE-L: 0.710 METEOR: 0.725
PathMem	Microscopic observation of the pathology slide reveals exuberant granulation tissue with characteristic fibroblastic proliferation and myxoid changes, typical for a high-grade myxofibrosarcoma. No residual sarcoma tissue is identified in the examined section, suggesting a primarily reactive tissue response. The tissue architecture is notably altered by the presence of foreign material, as evidenced by the defect in the suture material. Based on these observations, the diagnosis is myxofibrosarcoma, high grade.	0.821	0.675	BLEU-1: 0.728 BLEU-2: 0.683 BLEU-3: 0.628 BLEU-4: 0.600 ROUGE-L: 0.785 METEOR: 0.731
KG Activation	idx: 6599 soft tissue sarcoma → HAS_DIAGNOSTIC_CLUE → high-grade myxofibrosarcoma（a high-grade subtype diagnosis within soft tissue sarcoma） idx: 6599 soft tissue sarcoma → HAS_DIAGNOSTIC_CLUE → myxofibrosarcoma（specific histologic subtype identified in pathology）	Provides grading information for the **tumor's high malignancy** and aggressiveness. Provides specific histological subtype information for **soft tissue sarcoma**.		

Figure 9: Comparison of WSI-based and NLU-based evaluations. Green indicates agreement with the reference, red denotes deviations, orange marks missing ground-truth content, underlined text reflects template-style language, and blue highlights knowledge graph–grounded concepts.