TeamPath: Building MultiModal Pathology Experts with Reasoning AI Copilots

Anonymous ACL submission

Abstract

Advances in AI have introduced several strong models in computational pathology to usher it into the era of multi-modal diagnosis, analysis, and interpretation. However, the current pathology-specific visual language models still lack capacities in making diagnosis with rigorous reasoning paths as well as handling divergent tasks, and thus challenges of building AI Copilots for real scenarios still exist. Here we introduce TeamPath, an AI system powered by reinforcement learning and router-enhanced solutions based on large-scale histopathology multimodal datasets, to work as a virtual assistant for expert-level disease diagnosis, patch-level information summarization, and cross-modality generation to integrate transcriptomic information for the clinical usage. We also collaborate with pathologists from Yale School of Medicine to demonstrate that TeamPath can assist them in working more efficiently by identifying and correcting expert conclusions and reasoning paths. Overall, TeamPath can flexibly choose the best settings according to the needs, and serve as an innovative and reliable system for information communication across different modalities and experts.

1 Introduction

Pathological diagnosis is a complex yet essential component of clinical decision-making. In this manuscript, we present TeamPath, a framework that augments VLMs with multi-modal reasoning and a task-sensitive routing mechanism, enabling robust performance in several pathology-related tasks. Our approach begins with the careful selection of base models and the design of medical-specific prompts to curate high-quality, reasoning-enriched training data. Through comprehensive analyses, we demonstrate both the necessity of equipping VLMs with reasoning capabilities to address complex pathology tasks and the importance of constructing high-quality datasets for model success. We further showcase the effectiveness of TeamPath across diverse downstream applications, including multi-modal pathology visual question answering (Pathology VQA) and caption summarization. By leveraging an LLM-driven router, TeamPath dynamically selects the most suitable strategy to meet task requirements, functioning as a reliable and adaptive system. Importantly, we invite pathologists to evaluate the model's reasoning pathways, thereby validating its practical utility as a medical assistant. Finally, we introduce a new task, known as spatial transcriptomic profiles generation, to assess the cross-modality generative ability of TeamPath. Overall, TeamPath provides a new avenue for integrative analyses that combine molecular and histopathological signatures.

2 Methods

Problem definition. In this manuscript, we aim to construct a pathology-expert-level visual language model $\mathcal{M}()$ which accepts text prompts T and pathology image P as inputs. The outputs of our model follow the instructions and information provided in T and P. To train $\mathcal{M}()$, we collect a dataset $D_p = \{(T_1, P_1), ..., (T_n, P_n)\}_1^n$ with n items for training, and transfer the trained model to various downstream applications.

Constructing TeamPath as a system. To enhance our system's multitasking capabilities, we adopted a method commonly used in current basic model development, namely training a language model-based router ($\mathcal{R}()$) according to tasks and requirements. This router accepts questions as input data and outputs the model it selects to solve specific problems. The advantage of this design is to unify the TeamPath as a system for various downstream applications in digital pathology, and select the solution that best meets needs to save costs and improve model capabilities. We mark the best solution settings (one of the following choices:

Reinforcement Learning (RL) (Sheng et al., 2024), Supervised FineTuning (SFT) (Zheng et al., 2024), and test-time verification and correction (TTVC) inspired by test-time scaling (TTS) (Snell et al., 2025)) of each question, and train $\mathcal{R}$ with questions and choices. Here, RL is used for solving questions that require reasoning, and SFT is used for summarization and cross-modality generation. Since AI Copilot needs interactions with physicians, TTVC is used for tasks requiring human-AI collaboration. Current base model of TeamPath is Patho-R1-7B (Zhang et al., 2025), which is selected after carefully comparing it with different LMMs such as Qwen2.5VL-7B (Team, 2024), Qwen2.5VL-3B (Team, 2024), MedVLThinker-7B (Huang et al., 2025), PathGen-LLaVA-13B (Sun et al.), InternVL3-8B (Zhu et al., 2025), and MedGemma-4B (Sellergren et al., 2025).

Empowering TeamPath with reasoning capacities. To enhance the reasoning capabilities of TeamPath for complex pathological analysis, we adopt Group Relative Policy Optimization (GRPO), an efficient reinforcement learning algorithm that forgoes the critic model used in traditional PPO (Shao et al., 2024).

For each pathological query q, GRPO samples a group of G outputs $\{o_1, o_2, \ldots, o_G\}$ from the current policy π_θ and optimizes the GRPO objective, where the key novelty lies in the group-relative advantage estimation:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}. \quad (1)$$

Here, $\mathbf{r} = \{r_1, r_2, \ldots, r_G\}$ represents the reward scores for all outputs in the group, obtained from a reward model trained on the quality of pathological reasoning. This group-relative formulation eliminates the need for a separate value function V_ψ required in PPO, significantly reducing computational overhead while maintaining training stability.

For GRPO, the reward of question i is:

$$r_i = r(\hat{y}_i, y_i) = \begin{cases} 1, & \text{is\_equivalent } (\hat{y}_i, y_i) \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

where $\hat{y}$ and y represent the model outputs and observed answers, respectively. is_equivalent() is a function used to determine if the answer is correct or not.

In our ablation studies, we also consider introducing open-ended questions to model training; in that case, we utilize the BLEU score as a reward. The reward for closed-ended samples is the same, but for open-ended sample j, the reward is:

$$r_j = r(\hat{y}_j, y_j) = \text{BLEU}(\hat{y}_j, y_j). \quad (3)$$

The comparative nature of this approach aligns naturally with pathological diagnosis workflows, where medical experts simultaneously evaluate multiple diagnostic hypotheses. By learning from the relative quality of responses within each group, TeamPath develops more nuanced reasoning capabilities for tasks requiring differential diagnosis, evidence synthesis, and step-by-step pathological analysis.

In our ablation studies, we also consider Dynamic sAmpling Policy Optimization (DAPO) as an alternative reinforcement learning algorithm. DAPO removes the KL divergence and adjusts the group-level normalization method.

3 Results

Dataset and Method Overview. The curation of high-quality datasets is increasingly critical for advancing PFMs and VLMs, particularly in the era of multimodal reasoning and summarization. At the same time, careful attention must be paid to preventing data leakage to ensure unbiased evaluation of model performance. To this end, and leveraging prior data collection strategies, we distilled a subset of data from PathGen-1.6M (Sun et al.), which is a large-scale resource comprising nearly 10,000 WSIs and 1.6 million ROIs derived from TCGA data (Weinstein et al., 2013), for the usage in the finetuning stage with reinforcement learning. Reasoning data were constructed using COT templates generated based on the advanced reasoning model o4-mini (OpenAI, 2025), with subsequent quality validation performed by pathologists at Yale School of Medicine. Importantly, this dataset does not overlap with the benchmark testing set used for Pathology VQA evaluation (He et al., 2020b), namely PathMMU (Sun et al., 2024), which contains ROIs paired with questions across five diagnostic categories and represents one of the most advanced evaluation sources. In addition, another subset distilled from PathGen-1.6M was curated as the testing dataset for the ROI summarization task. To assess performance in cross-modality generation, we leveraged HEST-1K (Jaume et al., 2024) and STImage1K4M (Chen et al., 2024), two multi-omic histopathology collections to assess the

prediction of transcriptomic profiles as molecular signatures from ROIs. These two datasets are used to construct training, validation, and test sets. The overall data preprocessing workflow and sample sizes are summarized in Extended Data Figure 1.

The overall process of dataset curation and model training is summarized in Figures 1 (a)-(d). TeamPath emerges as a robust multimodal AI assistant for both disease analysis and modality generation. To refine its reasoning capabilities, we employ Group Relative Policy Optimization (GRPO) (Guo et al., 2025) to finetune the base model (the default setting is Patho-R1-7B), thereby enhancing its ability to perform reasoning over pathology images. With this capacity for structured reasoning, TeamPath demonstrates strong performance in addressing Pathology VQA tasks, as shown in our comprehensive benchmarking analysis. Importantly, the model also maintains high performance on tasks where reasoning is less critical, such as image summarization (known as caption generation) and cross-modality generation. This adaptability enables TeamPath to support task-specific optimization through either reinforcement learning or supervised finetuning. In collaboration with expert pathologists, we further demonstrate that TeamPath can function as a clinical copilot, assisting in applications such as correcting inaccurate conclusions and identifying flawed reasoning paths. Taken together, TeamPath advances both biomedical research and clinical practice in histopathology analysis. Finally, a comparative summary of task- and metric-specific rankings, shown in Figure 1 (e), demonstrates the superior performance of TeamPath across multiple dimensions.

TeamPath improves the performance of ROI-level assessment with reasoning ability. The increasing complexity of histopathology image analysis presents significant challenges for developing expert-level VLMs. One particularly demanding setting is Pathology VQA, which requires models to correctly respond to questions grounded in histopathology images. Unlike traditional classification tasks (e.g., disease-state classification or cancer cell identification), Pathology VQA involves a broader and more complex range of scenarios (He et al., 2020a) and demands higher accuracy in answer production. To evaluate model performances under this setting, we employ the recently published PathMMU dataset, which includes VQA pairs spanning five categories, ranging from expert-annotated questions to images from social media. Importantly, PathMMU is excluded from the training data of all evaluated models to ensure fairness. Reflecting the real-world requirements faced by pathologists, we emphasize the need for high-quality, fine-grained answers that integrate multimodal information and contribute meaningfully at the clinical level. Our baseline comparisons encompass (1) general-domain VLMs, including o4-mini, GPT-4o (Hurst et al., 2024), Qwen2.5VL-3B, Qwen2.5VL-7B (Team, 2024), and InternVL3-8B (Zhu et al., 2025); (2) medical-domain VLMs, including MedGemma-4B (Sellergren et al., 2025) and MedVLThinker-7B (Huang et al., 2025); (3) pathology-specific VLMs, including PathGen-LLaVA-13B (Sun et al.) and Patho-R1-7B (Zhang et al., 2025); and (4) a random-answer baseline. Model performance is assessed by computing accuracy relative to expert-generated answers within PathMMU, enabling a rigorous and fair benchmarking analysis. The comparison of sample size used for training and testing is shown in Extended Data Table 1, and we can see that the amount of images used in testing is large enough to support a general conclusion.

Figures 2 (a)-(c) show our benchmarking results across different categories, including PubMed, SocialPath, Atlas, EduContent, and PathCLS. PathMMU also pre-defines different sample types, and "overall" represents all testing samples in the selected category, "tiny_test" represents testing samples used for expert evaluation, and "test" represents the rest of the samples. We find that TeamPath outperforms all other baseline models, including domain-expert VLMs with similar or larger parameter size, such as Patho-R1-7B and PathGen-LLaVA-13B, in nearly all evaluations. TeamPath also performs better than strong general VLMs, such as o4-mini and GPT-4o, further demonstrating the strength of expert models in addressing medical challenges. Moreover, o4-mini and GPT-4o still perform better than most of the selected baselines, indicating they possess a certain level of understanding of pathological knowledge. Other general VLMs and medical VLMs performed poorly in this task. We further visualize the comprehensive benchmarking analysis, including ranking and accuracy of each method with all samples in Figure 2 (d), which shows that TeamPath also has the lowest rank by considering all categories jointly. Therefore, our experiment results show that introducing reasoning capacities to build pathology-expert VLMs can enhance their ability in making diagnoses, and thus

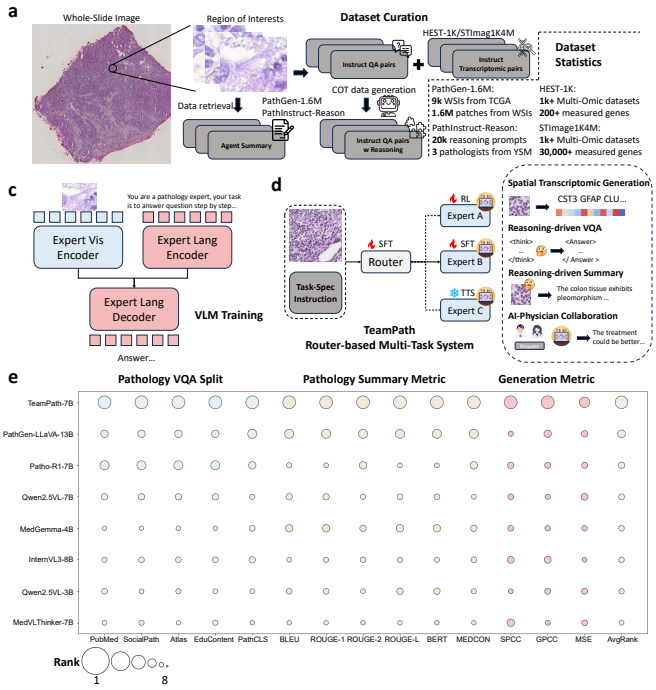


Figure 1: Landscape of TeamPath (a) Steps of dataset curation. We extract image-text pairs from a processed TCGA dataset (PathGen-1.6M). (b) Word cloud visualization of ROI captions (upper) and questions (bottom). (c) The core visual language model architecture of TeamPath. (d) TeamPath as a system with an LLM-enhanced router (with over 80% accuracy in choosing the correct approach) and the corresponding capacities in various downstream applications. The logo fire means that we need to adjust the parameters of models, and the logo snowflake means that we do not change the parameters. (e) Overall ranking list of different methods across tasks and metrics.

TeamPath can serve as a strong performer for the key feature identification and content understanding of ROIs.

To examine the performance of TeamPath specific in disease diagnosis, we utilize GPT-5.2 to extract the type of questions in PathMMU and select diagnosis-related questions to make comparison. According to Extended Data Figures 2 (a)-(b), TeamPath performs better than the second-best baseline Patho-R1-7B as well as random guessing in handling disease-diagnosis-related quires, and the improvement is consistent across most of the disease categories in both tiny group and large group. Therefore, TeamPath can also improve diagnostic accuracy after training.

To obtain a more intuitive understanding of the key contributions of TeamPath following reinforcement learning training, we selected two case studies where TeamPath provided the correct answer while other models failed to make accurate judgments.

Figure 3 highlights the importance of precise morphological criteria in recognizing lipoblasts. While several models incorrectly selected option C, describing large, clear vacuoles displacing the nucleus to the periphery, a hallmark of mature adipocytes. We found that TeamPath correctly identified option B as the defining feature of lipoblasts (Hisaoka, 2014). Lipoblasts are diagnostically recognized by the presence of moderately sized cytoplasmic fat vacuoles that indent or scallop the nucleus, a distinction that separates them from both mature adipocytes and other stromal features. By emphasizing nuclear indentation rather than displacement, TeamPath demonstrated accurate pathological reasoning aligned with standard diagnostic criteria. This correctness not only underscores the reliability of TeamPath in differentiating subtle histologic features but also highlights the critical nuance needed in distinguishing malignant lipoblastic cells from benign adipocytic processes. Moreover, Extended Data Figure 3 demonstrates that TeamPath correctly identified synaptophysin as the targeted marker in the immunohistochemical stain of section A. The brown, cytoplasmic staining pattern observed is a hallmark of synaptophysin, which is widely used as a marker of neuroendocrine differentiation. While other models misclassified the stain as estrogen receptor or S100 protein, TeamPath distinguished the subtle morphological and staining features that separate synaptophysin from

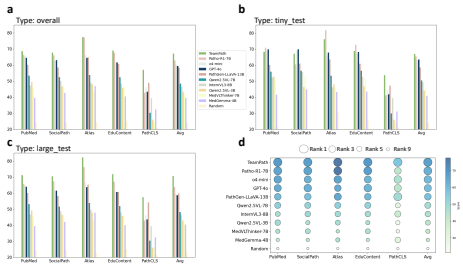


Figure 2: Benchmarking results with PathMMU for the pathology VQA task. We note that since we did not have information about the testing setting of PathGen-LLaVA-13B, we used results reported by the model creators in (Sun et al.). (a) Accuracy across different categories of all selected methods with all samples. (b) Accuracy across different categories of all selected methods with samples from a tiny set. (c) Accuracy across different categories of all selected methods with samples from a large set. (d) Joint visualization with accuracy and ranking information for all selected methods. The darker the bubble color, the higher the model score; The larger the bubble shape, the lower the model ranking.

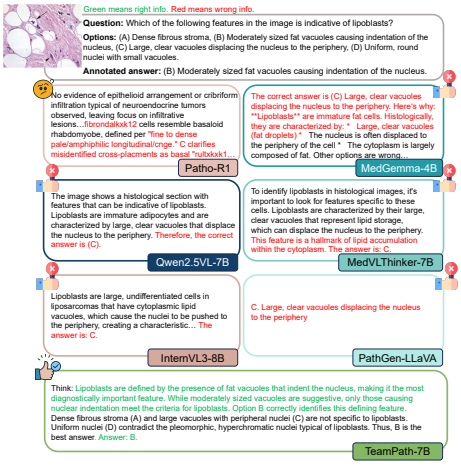


Figure 3: Case study (topic: synaptophysin, which is a precursor cell that develops into an adipocyte (fat cell)) based on the outputs from different models. We highlight the correct information with green text and incorrect information with red text. For the models with errors, we consider two cases. The first case is a wrong answer, and the second case is a confused reasoning path.

nuclear markers like estrogen receptor or more diffuse proteins such as S100. This highlights both the accuracy and interpretive strength of TeamPath in immunohistochemistry tasks, particularly in recognizing marker-specific staining patterns and avoiding common pitfalls that lead to misclassification. We also note that previous pathology expert models have obvious shortcomings, such as Patho-R1-7B's garbled output and PathGen-LLaVA's lack of interpretable diagnostic outputs. Instead, TeamPath can make correct identification supported by comprehensive explanations, explained in the information provided by the reasoning paths.

We also explored the contributions of different training strategies and highlighted the importance of selecting base models based on a set of ablation studies, discussed in Appendix G and Extended Data Figures 4 (a)-(d), as well as in Appendix H and Extended Data Figure 5 for the data ablation study.

TeamPath acts as a Copilot in the pathologists-AI collaboration system. Beyond demonstrating the capacity of TeamPath in handling VQA sets as a pathology expert, we further explore its potential as an AI-assisted collaborator (Mialon et al., 2023; Liu et al., 2025a). An effective copilot should not only provide accurate responses to user queries but also reduce the effort required to resolve them, thereby saving both time and cost. To this end, we designed an algorithm in TeamPath with test-time verification and correction (TTVC) (Snell et al., 2025; Liu et al., 2025a) and engaged expert pathologists from Yale School of Medicine (YSM) to collaborate with TeamPath in analyzing histopathology images and generating answers on demand. Specifically, we randomly subsampled 10 question–image pairs from each category within the PathMMU "tiny_test" set and examined two capacities: (1) the ability of TeamPath to act as an auto-verifier or auto-corrector for incorrect expert assessments, and (2) the ability of TeamPath to revise and correct reasoning pathways when human experts fail to provide accurate answers. When performing the TTVC, we prepared the inputs as the images, questions, and reasoning from experts, and performed verification and correction based on TeamPath. The overall paradigm for these two tasks is summarized in Figure 4 (a). Through this study, we aim to establish future paradigms of human–AI collaboration in biomedical research and clinical practice, highlighting the role of TeamPath as a reliable and strong copilot.

We jointly compared the expert-provided results with those corrected by TeamPath and visualized the corresponding accuracies in Figure 4 (b). Our analysis shows that TeamPath significantly improves accuracy across all PathMMU categories (p-value = 0.004), demonstrating that its corrective contribution is consistent and robust regardless of the source of pathology ROIs or questions. Notably,

even in categories where expert performance is relatively low, such as PubMed, TeamPath achieves substantial gains. These improvements demonstrate the effectiveness of TeamPath as a corrector, as reflected by the observed accuracy differences. To further illustrate this capability, we conducted a case study (Figure 4 (c)) in which the expert provided an incorrect answer, whereas TeamPath generated the correct response with an improved reasoning path. In this example, the task involved identifying characteristic nuclear features within the image. The expert's reasoning correctly accounted for cell size but overlooked nucleolar details, leading to an erroneous conclusion. In contrast, TeamPath integrated multiple features, including nuclear size, shape, staining depth, and prior knowledge of the cancer cell line, to eliminate incorrect options and arrive at the correct decision. Moreover, TeamPath was also able to revise flawed reasoning paths when experts could not provide an answer (e.g., "I do not know"), as shown in Extended Data Figure 6. In summary, through collaborative evaluation with pathologists, we demonstrate the capacity of TeamPath to not only fix erroneous answers but also provide explicit reasoning steps, thereby enhancing both the transparency and interpretability of model-assisted pathology analysis.

Here we provided more detailed analyses for understanding the case that TeamPath's discrepancies arising in correcting the feedback from different pathologists. We first compared the Pearson Correlation Coefficients (PCCs) among the three pathologists' accuracy. Extended Data Figure 7 (a) shows that the decision-making of our two pathologists is relatively consistent, while their accuracy rates on specific datasets also differ. Therefore, we can ensure a certain degree of diversity in expert selection used for human-AI collaboration experiment. Moreover, we also compared the number of corrected samples made by TeamPath for different pathologists, and Extended Data Figure 7 (b) shows that the number of corrections made by different experts is relatively similar across most datasets, with the sole exception being the SocialPath dataset. Since this dataset is primarily sourced from social media, the complex data pattern can pose challenges to our designed system. Our experiments demonstrate that TeamPath contributes to improving the performance across different pathologists, thereby exhibiting relatively broad applicability.

We have also performed ablation studies for the

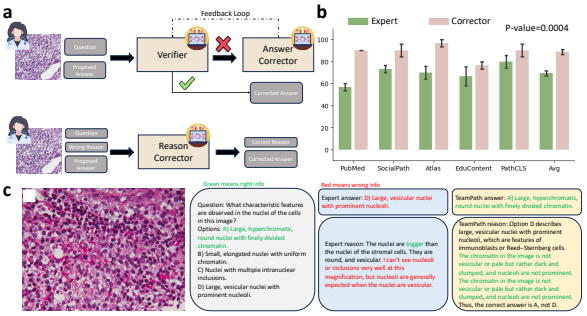


Figure 4: Results of using TeamPath as the answer corrector/reason corrector. TeamPath can work with pathologists together to improve the decision accuracy and provide explainable reasons to support the decision. (a) The illustration of self-verification/correction steps for both answers and reasoning paths. (b) Accuracy before and after correction based on selected samples from PathMMU. We report the average scores and standard deviation across three experts. The test is a one-sided Wilcoxon Rank-sum test. (c) A case study to demonstrate the power of TeamPath as an AI assistant.

verifier with three different choices (using the corrector, o3 (OpenAI, 2025), and o4-mini). Extended Data Figure 8 shows that using o4-mini can achieve the best performance on average, while it can also reduce the cost compared with using o3 or a more advanced model, and thus o4-mini is selected here to perform verification.

TeamPath performs better in summarizing the key information from histopathology images. In practical applications, pathology image analysis often extends beyond generating correct answers and reasoning steps to encompass the extraction of important image features for macroscopic or high-level descriptions. To evaluate this capability, we designed experiments aimed at summarizing histopathology information from different ROIs, thereby assessing the capacity of TeamPath to capture and convey high-level image content. For this purpose, we constructed a testing dataset by subsampling 3,000 images and their corresponding captions from PathGen-1.6M. In their setting, the original image caption is enhanced by multi-agent (GPT-4V as backbone) collaboration. These captions were further annotated to include tissue- and disease-state information based on prompting Deepseek-R1 (Guo et al., 2025) to extract the answer from the input caption. To support training, we curated a separate dataset of 50,000 images, ensuring no overlap with the testing set. For benchmarking, we employed the same set of baseline models used in the Pathology VQA experi-

ments. Model performance was evaluated using multiple similarity metrics between generated summaries and reference captions, including BLEU (Papineni et al., 2002), ROUGE-1/2/L (Lin, 2004), BERTScore (Zhang et al.), and MEDCON (Soldaini and Goharian, 2016). All metrics were scaled to a 0–100 range, with higher values indicating better performance.

Figure 5 (a) compares the performance of TeamPath with other VLMs across all selected metrics on the testing set. TeamPath consistently outperforms the baselines across every metric, demonstrating its strength in generating summaries that align closely with reference annotations in both content and structure. To provide a holistic assessment, we further visualized the aggregated rankings and average scores of all methods in Figure 5 (b), which highlights the leading performance of TeamPath across the joint set of evaluation metrics. Recognizing that performance may vary by sample source, we also examined model performance across specific tissue and disease contexts. We first analyze the produced categories in the testing dataset and illustrate the distribution of top 10 categories in Extended Data Figure 9. Among these labels, we select the top 1 label in each class, including adenocarcinoma and breast tissue, for further investigation. Figure 5 (c) reports ROUGE-L and BERT scores for samples from patients with adenocarcinoma, while Figure 5 (d) shows results for breast tissue samples. In both cases, TeamPath maintains superior performance compared with competing baselines. As an illustrative case study, Figure 5 (e) presents an example output from TeamPath, which accurately captures key organizational and pathological features—such as elongated spindle-shaped cells with eosinophilic cytoplasm and elongated nuclei, characteristic of smooth muscle cells. By contrast, outputs from baseline models (Extended Data Figure 10) contain less precise descriptions and, in some cases, incorrect content, further underscoring the advantages of TeamPath in summarization tasks.

Therefore, we conclude that TeamPath demonstrates as a strong performer in providing the high-level interpretations with pathology features of assigned ROIs.

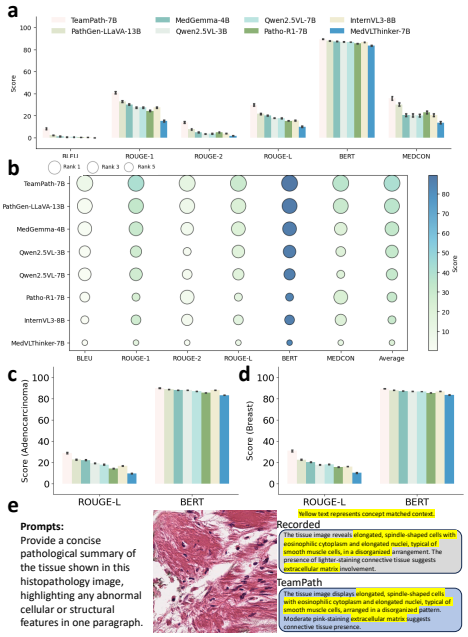


Figure 5: Benchmarking results of the caption summary task. (a) Performances of different methods for summarizing the caption based on ROI-level information across all metrics. We report the average scores and scaled standard deviation (0.1*sd) with all samples in the testing set. (b) Joint visualization with metric scores and ranking information for all selected methods. The darker the bubble color, the higher the model score; The larger the bubble shape, the lower the model ranking. (c) ROUGE-L and BERT scores based on samples from the selected disease across all methods. (d) ROUGE-L and BERT scores based on samples from the selected tissue across all methods. (e) A case study of caption summary generation based on TeamPath.

TeamPath introduces new modalities with a cross-modality generation pipeline. Building on our previous research and the existing literature, we observe that current histopathology image analyses primarily rely on textual and visual interpretations. However, given the breadth of biological signatures that can contribute to disease modeling and diagnosis, there is a clear opportunity to design new pipelines that integrate molecular information with histopathology features. Such integration can enable the generation of new modalities and provide deeper insights into cellular heterogeneity, lineage tracing, and disease mechanisms (Chen et al., 2025b; Song et al., 2024). Therefore, we finetune TeamPath using paired histopathology images and transcriptomic profiles generated with the Visium technology (Genomics), a platform for spatial transcriptomics (ST). Each ST spot includes

a histopathology image as background and a corresponding gene expression profile. Inspired by Cell2Sentence (Levine et al., 2024) and Loki (Chen et al., 2025b), we convert gene expression profiles into ranked gene lists, ordering genes from highest to lowest expression. The task is then to generate these "spot sentences" and map them back into the transcriptomic space. For training and evaluation, we use two of the largest public datasets: HEST-1K (invasive ductal carcinoma, IDC) and STImage1K4M (brain tissue). HEST-1K includes a broad range of cancer datasets, whereas STimage1K4M contains samples from both disease and normal tissues, thereby enhancing the modeling of ST data. Baseline models for this task include the same VLMs evaluated in the Pathology VQA setting, supplemented with Cell2Sentence-1B. Model performance is assessed using Spot-level Pearson Correlation Coefficient (SPCC), Gene-level Pearson Correlation Coefficient (GPCC), and mean squared error (MSE). For SPCC and GPCC, higher values indicate better performance, whereas lower MSE values reflect higher accuracy.

Figures 6 (a) and (b) demonstrate that TeamPath outperforms all baseline methods when evaluated by both SPCC and MSE across datasets from different sources, underscoring its ability to generate spot-level gene expression profiles that closely resemble measured results. Extended Data Figures 11 (a) and (b) further confirm TeamPath's better performance in GPCC, highlighting its capacity to preserve gene-level heterogeneity across spatial spots. To examine the impact of base model selection on cross-modality generation, we finetuned Qwen2.5VL-7B for the same task and compared it with TeamPath. As shown in Extended Data Figures 11 (c) and (d), TeamPath, which was built on a pathology-knowledge-enhanced VLM, outperformed the finetuned Qwen-series model. We also emphasize the importance of task-specific finetuning, supported by the clear performance gap between the unadapted base model and TeamPath in generating high-quality expression profiles. UMAP visualizations of the generated profiles (Figures 6 (c) and (d)) show that outputs from TeamPath are more structured and closely aligned with reference profiles compared to those from the base model. This observation is further validated by cluster-level heatmaps of gene expression patterns in brain (Figure 6 (e)) and IDC (Figure 6 (f)) datasets, where TeamPath more accurately recapitulates the biological signal present in the ground truth data. Collectively, these findings demonstrate that the effectiveness of TeamPath in cross-modality generation arises from both the choice of a pathology-informed base model and targeted task-specific finetuning. With these advantages, TeamPath represents a promising approach for generating in-silico or unseen expression profiles directly from histopathology images, thereby providing molecular-level insights into disease phenotypes.

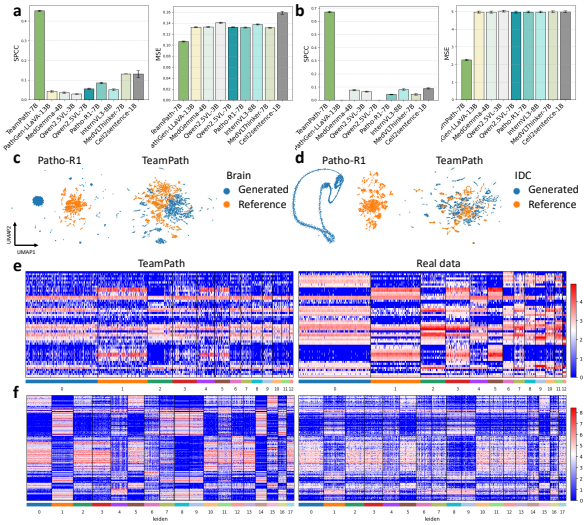


Figure 6: Evaluation of model performances for transcriptomic profile generation. (a) SPCC (higher is better) and MSE (lower is better) scores across different methods for the brain tissue. We report the average scores and scaled standard deviation (0.1*sd) for better visualization. (b) SPCC and MSE scores across different methods for the IDC samples. We report the average scores and scaled standard deviation (0.1*sd) for better visualization. (c) UMAP visualization from the testing set of brain to compare the generated results between Patho-R1 (base) and TeamPath colored by data sources. (d) UMAP visualization from the testing set of IDC to compare the generated results between Patho-R1 (base) and TeamPath colored by data sources. (e) Comparison of expression profiles between generated data and real data based on the brain tissue. (f) Comparison of expression profiles between generated data and real data based on the IDC samples.

4 Discussion

Here we present TeamPath, an advanced AI copilot to advance research in computational pathology and disease diagnosis by formulating a multi-task AI assistant, which supports various tasks with an automatic router for solution selection. TeamPath demonstrates as a strong method in working with physicians in addressing key pathology challenges.

References

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025a. SFT or RL? an early investigation into training r1-like reasoning large vision-language models. *Transactions on Machine Learning Research*.

Jiawen Chen, Muqing Zhou, Wenrong Wu, Jinwei Zhang, Yun Li, and Didong Li. 2024. Stimage-1k4m: A histopathology image-gene expression dataset for spatial transcriptomics. *Advances in Neural Information Processing Systems*, 37:35796–35823.

Weiqing Chen, Pengzhi Zhang, Tu N Tran, Yiwei Xiao, Shengyu Li, Vrutant V Shah, Hao Cheng, Kristopher W Brannan, Keith Youker, Li Lai, and 1 others. 2025b. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, pages 1–15.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*.

10X Genomics. Visium technology.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020a. Pathological visual question answering. *arXiv preprint arXiv:2010.12435*.

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

Masanori Hisaoka. 2014. Lipoblast: morphologic features and diagnostic value. *Journal of UOEH*, 36(2):115–121.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025. Medvlthinker: Simple baselines for multimodal medical reasoning. *arXiv preprint arXiv:2508.02669*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis Langlotz, and 1 others. Radgraph: Extracting clinical entities and relations from radiology reports.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro Pérez, Sophia J Wagner, Anurag Jayant Vaidya, Richard J. Chen, Drew FK Williamson, Ahrong Kim, and Faisal Mahmood. 2024. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Daniel Levine, Syed A Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, and 1 others. 2024. Cell2sentence: Teaching large language models the language of biology. In *International Conference on Machine Learning*, pages 27299–27325. PMLR.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries.

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

Tianyu Liu, Simeng Han, Xiao Luo, Hanchen Wang, Pan Lu, Biqing Zhu, Yuge Wang, Keyi Li, Jiapeng Chen, Rihao Qu, and 1 others. 2025a. Towards artificial intelligence research assistant for expert-involved learning. *arXiv preprint arXiv:2505.04638*.

Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, and 1 others. 2025b. Part i: Tricks or traps? a deep dive into rl for llm reasoning. *arXiv preprint arXiv:2508.08221*.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, and 1 others. 2024. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.

OpenAI. 2025. Openai o3 and o4-mini system card.

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

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction.

Andrew H Song, Mane Williams, Drew FK Williamson, Sarah SL Chow, Guillaume Jaume, Gan Gao, Andrew Zhang, Bowen Chen, Alexander S Baras, Robert Serafin, and 1 others. 2024. Analysis of 3d pathology samples using weakly supervised ai. *Cell*, 187(10):2502–2520.

Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, and 1 others. 2024. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, pages 56–73. Springer.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Jingxiong Li, Xuan Gong, XINHENG LYU, Tao Lin, and 1 others. Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In *The Thirteenth International Conference on Learning Representations*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, and 1 others. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, and 1 others. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Weixun Wang, Shaopan Xiong, Gengru Chen, Wei Gao, Sheng Guo, Yancheng He, Ju Huang, Jiaheng Liu, Zhendong Li, Xiaoyang Li, and 1 others. 2025. Reinforcement learning optimization for large-scale learning: An efficient and user-friendly scaling library. *arXiv preprint arXiv:2506.06122*.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

Wenchuan Zhang, Penghao Zhang, Jingru Guo, Tao Cheng, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. 2025. Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner. *arXiv preprint arXiv:2505.11404*.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. 2026. On-policy RL meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. In *The Fourteenth International Conference on Learning Representations*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Method Details

Ablation studies of training framework. To demonstrate the efficiency and optimization of our training framework for this system, we have considered several training strategies, including 1. Supervised finetuning (SFT), which collects paired data with images and queries as inputs and answers as outputs; 2. Reinforcement Learning (RL), which utilizes the same input and output data, but we train TeamPath with GRPO (Shao et al., 2024) or DAPO (Yu et al., 2025) to tackle the reasoning capacity of selected base models; 3. SFT+RL, which utilizes the paired data with images, queries, and reasoning paths as inputs and answers as outputs, to train TeamPath with SFT and then with RL. The first step of SFT training ensures the model acquires knowledge in relevant pathological domains, while the second step of RL training enhances the model's generalization capabilities. Our base models used for ablation studies include Qwen2.5VL-7B and Patho-R1-7B.

Using TeamPath as an AI Copilot to help pathologists. To formalize our method as an AI Assistant, we consider two case studies inspired by the Path VQA experiments with pathologists. We invite the pathologists to answer 50 questions extracted from PathMMU from the five categories, and record their answers as well as reasoning steps. Our first case is a verifier-corrector pipeline, which can detect the incorrect answers made by pathologists and generate the correct answers. Our pipeline utilizes one verifier (a VLM, default as o4-mini) to verify whether the answers and questions proposed by pathologists are correct or not. If it is justified as wrong, we will call the corrector (also a VLM, default as TeamPath used for Pathology VQA) to fix it. Otherwise, the correct answer will be returned. We have a specific threshold to limit the number of epochs in this loop. Our algorithm is summarized in Algorithm 1. We define the success of a self-verification/correction system as follows: if the expert answer is correct, or the expert answer is wrong but the answer produced by this system is correct.

The second case is a reasoning-correction pipeline. Here we have a reasoning corrector, which takes the wrong answers and reasoning paths from pathologists, and generates the correct reasoning path with the correct answer. To improve efficiency, we consider an offline version, that is, by collecting pathologists' reasons and answers first, we then utilize TeamPath to check or correct these questions. This pipeline can detect the wrong information provided in the reasoning process and generate the correct thinking steps. These two components focus on different aspects and work together as a prototype for building an AI Copilot that can work with pathologists and be deployed in the medical system. We have provided an example of correction in the main text. Our prompts used in these two pipelines are summarized in Appendix F.

Algorithm 1 Verifier-Corrector Pipeline in TeamPath.

Input: Question Q_M, pathology image I_S, human answer O_A, reasoning path O_R, verify prompt T_V, correct prompt T_C, number of iteration N.
Helper Models: Verifier $\mathcal{M}_v$ (An advanced LMM, such as O4-mini), corrector $\mathcal{M}_c$ (An pathology-specific LMM, such as TeamPath with RL finetuning), concatenation function $\cdot||\cdot$.
Output: Corrected outputs O_C

INIT: initialize all parameters.
if $\mathcal{M}_v(T_V, Q_M||O_R||O_A, I_S)$ is True **then**
    $O_C = O_A$
    Return O_C
end if
for i in N steps **do**
    $O_i, R_i = \mathcal{M}_c(T_C, Q_M||O_R||O_A, I_S)$
    if $\mathcal{M}_v(T_V, Q_M||R_i||O_i, I_S)$ is True **then**
        $O_C = O_i$
        Return O_C
    else
        $O_R = R_i$
        $O_A = O_i$
    end if
end for
$O_C = O_A$
Return O_C

Adapting TeamPath for image summarization and cross-modality generation. To summarize the concepts in histopathology images and further generate image caption, we finetune our base model based on the paired image-caption dataset with the corresponding training set, and we also prepare 3000 samples which are only used for testing, and we utilize Deepseek-R1 (Guo et al., 2025) to extract the disease state and tissue source of the testing samples based on their captions. Our finetuning step follows the setting in Instruction-Tuning imple-

mented in Llama-factory. We construct 10 different prompts to ask TeamPath for generating the image captions to reduce the bias of prompt information in the training process.

To perform the cross-modality generation task, we also finetune our base model based on the paired image-transcriptomic profile dataset with the corresponding training datasets. We select sequence data that comes from different batches and resources, but the same tissue/disease, to build a testing dataset. To transfer the information in gene expression space to text space, we first rank the genes of each spot based on their expression profiles and select the top 100 genes to formulate them in natural language. We then train a linear regressor that takes the natural language information as inputs and original gene expression profiles as outputs based on the training dataset, which finally gives us a method to decode the language information back to gene expression levels. We finetune our base model with the same approach used in image summarization and also construct 10 different prompts to ask TeamPath for generating gene expression profiles.

The prompts used in this section can also be found in Appendix F.

Evaluations. In this manuscript, we consider task-specific evaluation (Pedregosa et al., 2011; Virtanen et al., 2020) and follow the settings from previous works with shared tasks.

For the evaluation of Path VQA and Human-AI collaboration tasks, we utilize accuracy as a metric. The generated answer should be precisely matched with the provided answer. A higher accuracy represents a better method.

For the evaluation of the image caption summarization task, we utilize several metrics that can measure the similarity between the generated text and the provided text. These metrics include BLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERT score, and MEDCON (Papineni et al., 2002; Lin, 2004; Zhang et al.; Jain et al.; Yim et al., 2023), supported by a recent publication (Van Veen et al., 2024). We also consider the average score across these metrics. Here are the descriptions:

- BLEU: The BiLingual Evaluation Understudy (BLEU) score evaluates the quality of generated text by breaking both the generated output and the reference text into n-grams, then comparing the overlap between the two sets. The score ranges from 0 to 1 and is typically scaled to a range of 0 to 100, with higher values indicating better model performance.

- ROUGE: The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score assesses text quality by computing the F1 score from n-gram overlaps between the generated text and the reference text. In this framework, n-grams from the generated text are treated as predictions, while those from the reference text serve as labels. Precision, recall, and the F1 score are calculated using the counts of matching n-grams and their lengths. ROUGE-1 measures unigram overlap, ROUGE-2 measures bigram overlap, and ROUGE-L measures the longest common subsequence. The score ranges from 0 to 1 and is typically scaled to a range of 0 to 100, with higher values indicating better model performance.

- BERT: The Bidirectional Encoder Representations from Transformers (BERT) model is pre-trained on large-scale text corpora for language understanding and excels at producing rich text representations. The BERTScore metric leverages this capability by measuring the similarity between embeddings of the generated text and the reference text. The score ranges from 0 to 1 and is typically scaled to a range of 0 to 100, with higher values indicating better model performance.

- MEDCON: MEDCON limits the recognized concepts and entities to the semantic groups defined in QuickUMLS (Soldaini and Goharian, 2016), including Anatomy, Chemicals, Drugs, Device, Disorders, Genes, Molecular Sequences, Phenomena, and Physiology. These concepts are extracted from both the generated text and the reference text, and the F1 score is calculated based on the overlap between the two sets. The score ranges from 0 to 1 and is typically scaled to a range of 0 to 100, with higher values indicating better model performance.

A higher score of these metrics represents a better method.

For the evaluation of the cross-modality generation task, we consider spot-level Pearson Correlation Coefficient (SPCC), gene-level PCC (GPCC), and Mean Squared Error (MSE) as metrics. Higher SPCC and GPCC scores represent a better method,

while a lower MSE score represents a better method. These metrics are computed between generated gene expression profiles and observed gene expression profiles.

Baselines. Our baseline methods cover current state-of-the-art (SOTA) open-source LMMs based on the open source movement in scientific research and the powerful influence of open source models. Moreover, there are a few powerful closed-source models focusing on digital histopathology. We apply the access to PathChat (Lu et al., 2024) but have not received the authorization. These models include MedGemma-4B, Qwen2.5VL-3B, Qwen2.5VL-7B, MedVLThinker-7B, InternVL3-8B, PathGen-LLaVA-13B, and Patho-R1 (7B). MedGemma-4B is an open-source VLM released by Google based on finetuning Gemma with multimodal medical data. Qwen2.5VL-3B and Qwen2.5VL-7B are open-source VLMs from the Qwen team, Alibaba Cloud. They are trained with multimodal data in the general domain. InternVL3-8B is an open-source VLM released by OpenGVLab, and it is also trained with multimodal data from the general domain. For pathology-specific models, we consider PathGen-LLaVA-13B (Sun et al.), which is finetuned based on LLAVA 13B (Liu et al., 2023) with instruction data from PathGen; as well as Patho-R1, which has a pathology-specific image encoder and is finetuned based on Qwen2.5VL-3B with reasoning data.

For the cross-modality generation task, we also consider a task-specific baseline method, known as Cell2Sentence (1B) (Levine et al., 2024). Cell2Sentence is finetuned with instructions and single-cell transcriptomic data from atlas-level datasets based on Pythia. This model can generate cells based on instructions.

B Code and Data Availability

We utilize NCSA, YCRC, and TokyoU HPC platforms to perform experiments. To train TeamPath, we utilize 32 NVIDIA H100 cores and 8 NVIDIA H200 cores for 24 hours. The CPU memory upper bound is 80GB. The codes can be found in `https://github.com/HelloWorldLTY/TeamPath`, and the license is the MIT license.

The information on the datasets used in this manuscript can be found in Supplementary File 1. To access TCGA data, an authorized account is required. To protect personal privacy, we will not release experts' answers.

C Acknowledgments

We thank Mr. Tong Ding for his suggestion on model training and task selection.

D Author Contributions

T.L. and W.X. designed this study. T.L., W.X., and H.Q. ran all the experiments. H.W., P.H., M.D. performed human evaluation. All authors involved in writing and reviewing. H.Z. supervised this study.

E Institutional Review Board (IRB) Approval.

This project has received approval from Yale IRB, with project number 2000039055.

F Prompt list

The prompt used for the Pathology VQA is:

Your task: 1. Think through the question step by step, enclose your reasoning process in <think>...</think> tags. 2. Then provide the correct single-letter choice (A, B, C, D,...) inside <answer>...</answer> tags. 3. No extra information or text outside of these tags.

The prompt used for the self-verifier is:

You are an expert in pathology. You are given a QUESTION and a PROPOSED SOLUTION. Your job is to: 1. Break down each component of the proposed solution. 2. Think step by step to verify if the proposed solution is correct given the question and the figure. 3. Write a line of the form "The proposed solution is correct" or "The proposed solution is incorrect" at the end of your response based on your analysis. QUESTION: question. PROPOSED SOLUTION: solution.

The prompt used for the self-corrector is:

You are also given a question and an analysis for the question. Your job is to outline your step-by-step thought process for deriving a correct solution and also write down the correct solution. Using this format: <think> Your step-by-step reasoning of the question and solution </think><answer> Your final answer </answer> Question: question Solution: out_verifier.

The prompt used for the reason corrector is:

You are given QUESTION, REASON, and SOLUTION. Your task is to correct the REASON and SOLUTION. QUESTION: question. SOLUTION: solution. REASON: reason The REASON is WRONG. Your solution:

The prompts used for training TeamPath for caption summary include:

Provide a concise pathological summary of the tissue shown in this histopathology image, highlighting any abnormal cellular or structural features in one paragraph. Based on the visual characteristics in this image, summarize the likely histological diagnosis and key indicators leading to it in one paragraph. Describe the main histopathological patterns visible in this image and summarize what they suggest about the tissue state in one paragraph. Summarize the key morphological findings in this histopathology image, including any signs of malignancy, inflammation, or necrosis in one paragraph. Generate a pathology report-style summary based solely on this histological section, mentioning tissue type, grade, and diagnostic clues in one paragraph. Briefly summarize the clinical implications of the abnormalities visible in this histopathology image in one paragraph. From this histopathology image, extract and summarize the most diagnostically relevant features in one paragraph. Identify and summarize any histopathological hallmarks (e.g., mitotic figures, glandular formation, stromal invasion) present in the image in one paragraph. Write a summary suitable for a pathology trainee explaining what this histopathology image represents and why in one paragraph. Provide an expert-level summary of the pathological findings in this histopathology image, including your confidence in the assessment in one paragraph.

The prompts used for training TeamPath for cross-modality generation (using IDC as an example) include:

Generate a list of 100 genes in order of descending expression from one spot shown in the histopathology image in IDC disease. Cell sentence:, Produce a list of 100 gene names in descending order of expression which represent the expressed genes from one spot shown in the histopathology image in IDC disease. Cell sentence:, Create a ranked list of 100 genes in decreasing order of expression from one spot shown in the histopathology image in IDC disease. Cell sentence:, List the top 100 expressed genes from one spot shown in the histopathology image in IDC disease. Cell sentence:, Identify the highest expressed 100 genes in decreasing order of expression from one spot shown in the histopathology image in IDC disease. Cell sentence:, Enumerate a list of 100 genes in descending order of expression from one spot shown in the histopathology image in IDC disease. Cell sentence:, Compile a descending order list of 100 expressed genes from one spot shown in the histopathology image in IDC disease. Cell sentence:, Present a sequence of 100 genes ordered by decreasing expression level from one spot shown in the histopathology image in IDC disease. Cell sentence:, Generate an ordered list of 100 genes by decreasing expression level from one spot shown in the histopathology image in IDC disease. Cell sentence:, Assemble a list of 100 genes from highest to lowest expression from one spot shown in the histopathology image in IDC disease. Cell sentence:

G SFT vs RL: Comparison for training strategies.

How to train a mature reasoning model has always been a controversial topic, as there exist several different strategies and their performances and ranks might vary under different task settings or experiment settings (Chu et al., 2025; Wang et al., 2025; Liu et al., 2025b). Meanwhile, this kind of discussion has not been investigated in training a large reasoning model for histopathology analysis, and thus, we consider several different approaches to provide an empirical analysis to select and interpret the best combination, which might inspire future directions or different researchers.

We conduct these experiments based on different base models as well as training strategies. Our base models include Qwen2.5VL-7B, which does not contain domain-specific knowledge and Patho-R1-7B, which contains domain-specific knowledge. We also consider different training strategies, including RL (GRPO), RL (DAPO), SFT, and SFT+RL (GRPO). Extended Data Figures 4 (a) and (b) show that GRPO can achieve a more obvious score improvement while DAPO cannot make an improvement, which implies that a mixture of tricks does not contribute to training a multi-modal pathology reasoning model. Extended Data Figures 4 (c) and (d) show that it is important to select a base model with domain knowledge for training, and performing SFT+RL or direct SFT settings does not benefit this task. Therefore, our optimal

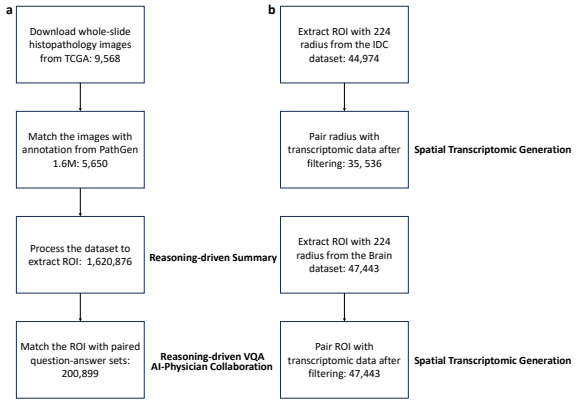
choice to build TeamPath for pathology VQA is Patho-R1-7B+RL (GRPO).

In conclusion, our results showcase that it is important to select a good model with domain knowledge to perform training, and for LMMs that can already possess domain knowledge, directly training them using RL policies can enhance their generalization capabilities without requiring specialized SFT training (also known as cold-start training). Our important findings also align with relevant research across different fields (Chen et al., 2025a; Zhang et al., 2026), which indirectly validates the reliability of our conclusions.

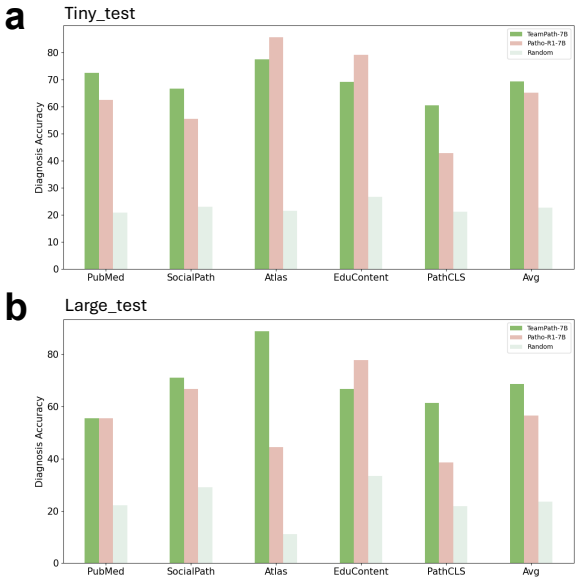
H Studies for training data ablation.

We also investigate if incorporating more diverse data could help TeamPath for generating a better reasoning path or not. To examine it, we compare the results between only using multi-choice VQA (mc VQA) data and using both multi-choice VQA and open-ended VQA (full VQA) data. Based on our evaluation results shown in Extended Data Figure 5 with validation from the PathMMU dataset, using full VQA data cannot boost TeamPath's performances in generalizing the results for solving questions in the PathMMU dataset. Therefore, our optimal setting only takes mc VQA for RL training. Details of the reward design for these two different types of data are explained in the Method section.

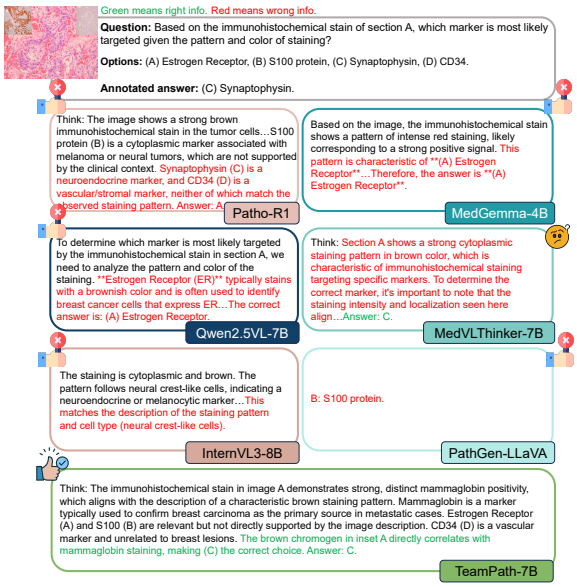
I Supplementary figures



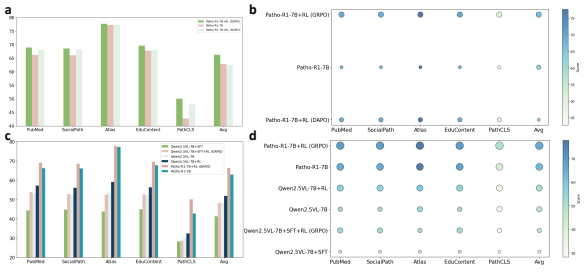
Extended Data Fig. 1: A flowchart of data-preprocessing used to train TeamPath.



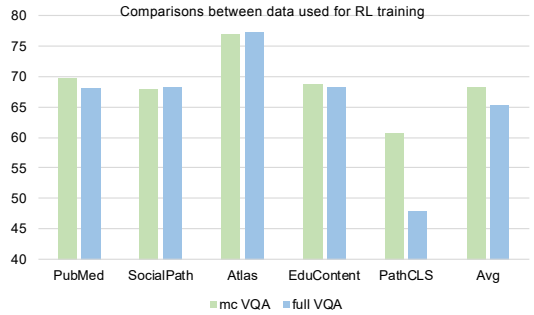
Extended Data Fig. 2: Accuracy of diagnosis-related questions in PathMMU. (a) Results reported based on the tiny_test dataset. (b) Results reported based on the large_test dataset.



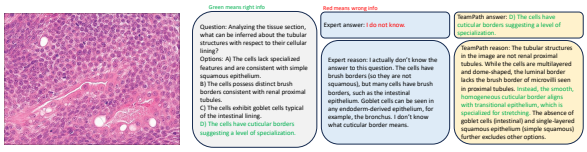
Extended Data Fig. 3: Case study (topic: lipoblast, which is a protein found in the presynaptic vesicles of neurons and neuroendocrine cells that plays a role in synaptic transmission) based on the outputs from different models. We highlight the correct information with green text and incorrect information with red text. For the models with errors, we consider two cases. The first case is a wrong answer, the second case is a confused reasoning path.



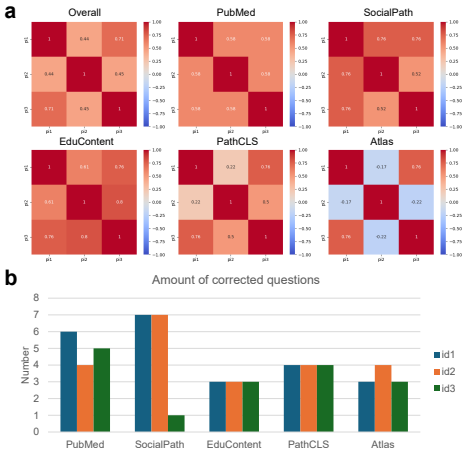
Extended Data Fig. 4: Training strategies optimization with different settings. The results are evaluated based on the PathMMU dataset. (a) Accuracy across different categories based on the base model and different RL strategies. (b) Accuracy and rank across different categories based on the base model and different RL strategies. (c) Accuracy across different categories based on different base models and different RL/SFT strategies. (d) Accuracy and rank across different categories based on different base models and different RL/SFT strategies.



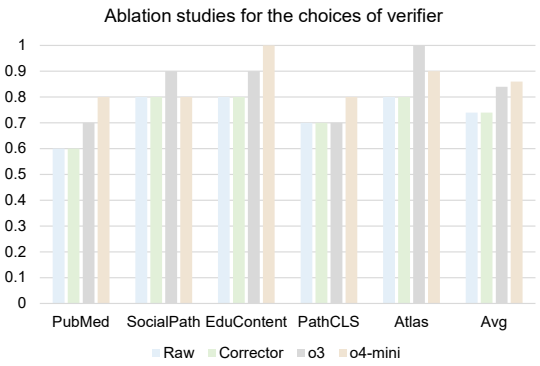
Extended Data Fig. 5: Comparisons for the PathMMU VQA question set with different training data. The metric is accuracy.



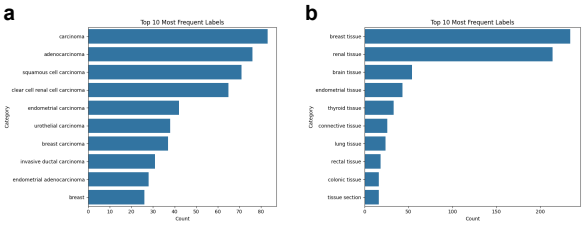
Extended Data Fig. 6: An example of using TeamPath to correct the reasoning path from pathologists.



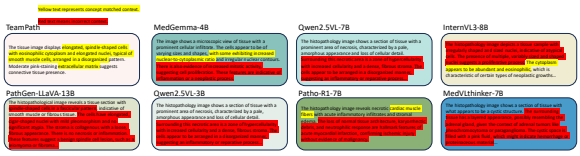
Extended Data Fig. 7: Results of expert feedback and TeamPath-corrected information. (a) The PCCs based on the accuracy of pathologists across different question categories in PathMMU. (b) The number of corrected samples for each pathologist made by TeamPath.



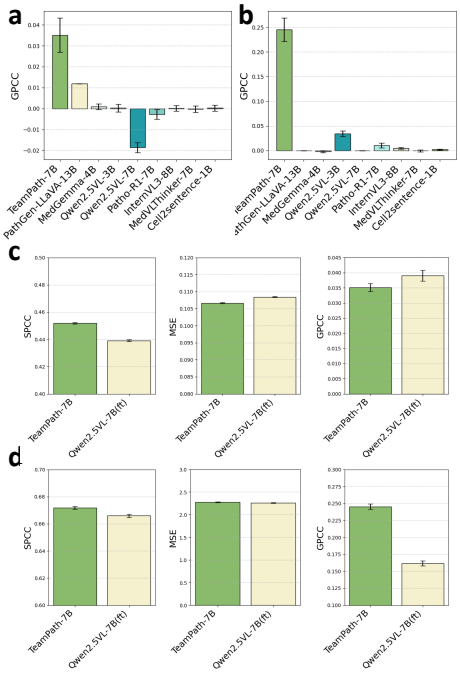
Extended Data Fig. 8: Ablation studies of using different models as verifiers.



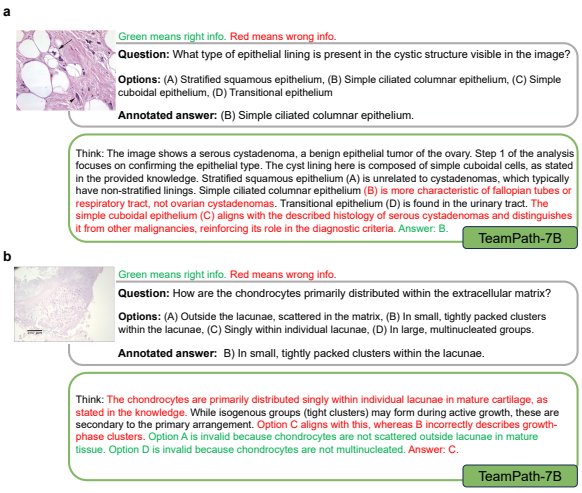
Extended Data Fig. 9: Distribution of image categories used for the testing of summarization performance. (a) Top 10 disease categories and their number. (b) Top 10 tissue categories and their number.



Extended Data Fig. 10: Examples of model outputs for caption summary tasks. We highlight both correct and incorrect information.



Extended Data Fig. 11: Extended analyses of spatial transcriptomic generation. (a) GPCC across different methods based on the brain tissue. (b) GPCC across different methods based on the IDC sample. (c) Comparison between finetuned Qwen2.5VL-7B and TeamPath based on the brain tissue. (d) Comparison between finetuned Qwen2.5VL-7B and TeamPath based on the IDC sample.



Extended Data Fig. 12: Case studies of TeamPath output for pathology VQA with (a) an incorrect reasoning path but a correct answer and (b) an incorrect reasoning leads to an incorrect answer.

J Supplementary Tables

Extended Data Tab. 1: Dataset statistics for training and testing sets.

Data type	Number of samples
Training_size (only MCA)	14,288
PubMed_test_tiny	281
PubMed_test	2,787
SocialPath_test_tiny	216
SocialPath_test	968
Atlas_test_tiny	208
Atlas_test	799
EduContent_test_tiny	255
EduContent_test	1,683
PathCLS_test	1,632
PathCLS_test_tiny	177
Test_total	9,006