

Enriching Context for Pathological Diagnosis via Multimodal Auxiliary Information

Anonymous ACL submission

Abstract

Although computational pathology has substantially advanced the automated analysis of pathological images, its reliance on visual features overlooks the multimodal context that human pathologists integrate, thereby constraining diagnostic accuracy. This study explores enhancing pathological diagnosis by providing models with three types of auxiliary information during inference, including clinical history, terminology explanations, and visual in-context examples. We fine-tune a vision-language model for pathological diagnosis with large-scale pre-training and instruction following data. Experiments across slide-level diagnosis, region of interest subtyping, and invasion detection tasks demonstrate significant improvements with enriched context. Our findings highlight the potential of enriching context with auxiliary information to bridge the gap between human diagnosis and computational pathology.

1 Introduction

The progress of artificial intelligence in the field of healthcare has revolutionized pathological diagnosis. Large-scale models (Lu et al., 2021a; Tang et al., 2023; Lin et al., 2023; Chen et al., 2024a; Xu et al., 2024) are enabled to perform automated diagnosis with remarkable efficiency, achieving performance comparable to or even surpassing that of human experts.

Despite these advancements, existing research focuses solely on exploiting visual features derived from pathological images. However, human diagnostic reasoning extends beyond visual patterns. Pathologists integrate clinical history when interpreting slides, including but not limited to patients’ age, gender, prior diagnosis and biological specimens. Moreover, the limited availability of pathological data fundamentally restricts the capabilities of vision-language models (VLMs) to identify diseases, leading to significant performance degradation when models encounter novel disease types.

These two constraints have built an information gap between human diagnosis and computational pathology. Therefore, it is important to enhance the inference phase by augmenting the textual context related to the patient and the disease. Furthermore, inspired by in-context learning, there is potential in integrating image examples during inference. Images can convey subtle details that cannot always be expressed through texts alone, offering richer and multi-dimensional contextual information that could improve diagnostic accuracy.

In this study, we evaluate the effect of enriching context with three types of auxiliary information in the pathological diagnostic tasks. Specifically, we investigate whether integrating **clinical history, terminology explanations and visual in-context examples** during inference can enhance the performance of VLMs. To validate the hypothesis, we conduct a series of experiments on Qwen2-VL (Wang et al., 2024a), which is further optimized for pathological diagnosis through fine-tuning. Notably, our findings provide compelling evidence that incorporating auxiliary information leads to apparent improvements in models’ performance, highlighting its significance in computational pathology. In summary, our contributions include:

1. We introduce a context-enriching strategy during inference that integrates (i) clinical history, (ii) terminology explanations, and (iii) visual in-context examples into the dialogue with large VLMs.
2. We develop a domain-adapted model of Qwen2-VL-7B-Instruct through fine-tuning, which is tailored to address the unique challenges in computational pathology.
3. Experiment results validate that our context-enriching method guides VLMs towards more accurate pathological diagnosis.

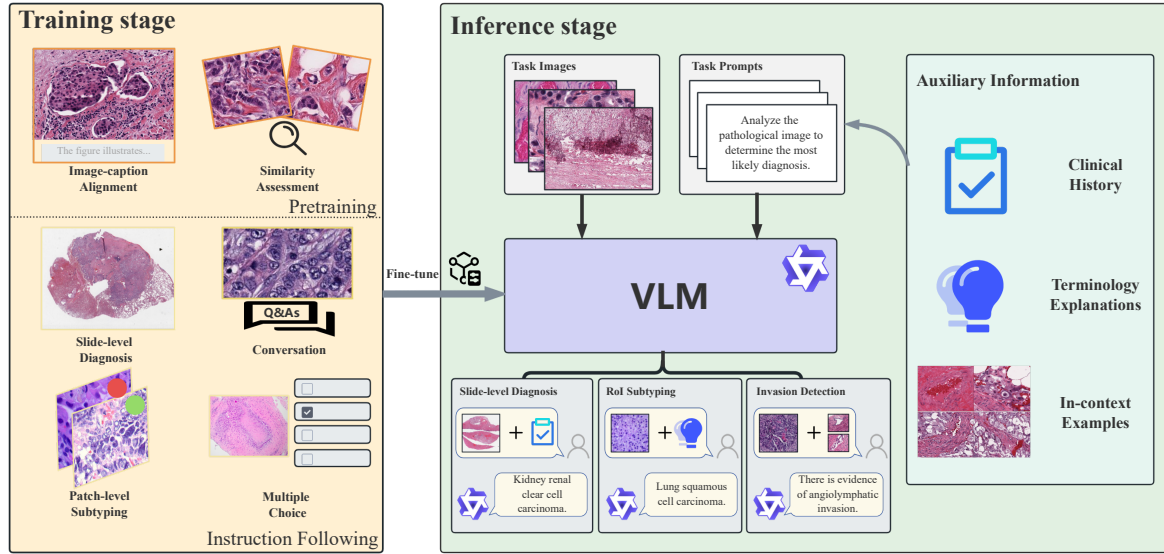


Figure 1: An overview of our framework. Left: the training data for fine-tuning. Right: integrating three types of auxiliary information during inference in three diagnosis-related tasks respectively.

2 Related Work

2.1 Pathological VLM

Recent advancements in computational pathology have facilitated the development of VLMs specially tailored for pathological image analysis. Early contributions include PLIP (Huang et al., 2023) and CONCH (Lu et al., 2024a), which establish foundational frameworks in this domain. Building upon these, PathChat (Lu et al., 2024b) emerges as a comprehensive vision-language generative assistant for pathology. Subsequently, SlideChat (Chen et al., 2024b) is introduced as the first VLM targeting gigapixel slide image analysis. CPath-Omni (Sun et al., 2024a) further advances the field by unifying both patch-level and whole-slide image analysis.

2.2 Enriching Context

Enriching context during inference is an effective strategy to adapt models for specialized domains, which augments the models with auxiliary information including task explanations (Recchia, 2021; Lampinen et al., 2022) and domain knowledge (Jin et al., 2023; Sural et al., 2024). In the medical field, Zakka et al. (2024) explores the retrieval of medical guidelines to generate reliable clinical responses. Wang et al. (2024b) integrates authoritative medical textbooks into the language models’ framework for medical question answering (QA). Neupane et al. (2024) proposes integrating patients’ medical

records to generate personalized responses using language models. However, the effect of context enrichment on the interpretation of pathological images remains insufficiently explored.

2.3 In-context Learning

Large-scale models have revealed a compelling learning strategy, in-context learning (ICL). Pioneered by GPT-3 (Brown et al., 2020), this paradigm enables models to learn from examples with labels during inference without any gradient updates. Extensive studies confirm that ICL is notably powerful with multiple benefits (Dong et al., 2022), including offering an interpretable framework for model interaction (Brown et al., 2020; Liu et al., 2021; Lu et al., 2021b; Wu et al., 2023), simulating human decision-making processes by learning from analogy (Winston, 1980) and reducing the computational costs for domain-specific adaptation.

In the field of pathology, ICL is still a novel approach with limited existing work. Nori et al. (2023) proposes Medprompt to identify the most relevant few-shot examples of medical QA pairs. Ferber et al. (2024) extends this framework to multimodal diagnostic applications in pathology. Liu et al. (2025) demonstrates the potential of ICL for improving pathological report generation. Nonetheless, prior literature has not yet investigated the potential of visual examples alone in enhancing the performance of VLM in pathological diagnosis.

3 Methods

3.1 Fine-tuning VLM

We fine-tune a foundational VLM for pathology-specific tasks. For parameter-efficient optimization, we employ a strategy that combines Low-Rank Adaptation (LoRA) (Hu et al., 2022) with Supervised Fine-Tuning (SFT) methodologies.

Our training data can be classified into two parts. The pretraining data enables the model to develop abilities of visual-text alignment and image-based feature exploration through exposure to large-scale pathological multimodal data. The instruction following data enables the model to handle diverse downstream applications more effectively, characterized by moderate scale and increased task complexity. The detailed data summarization is listed in Section 4.

Pretraining Data General-purpose VLMs may struggle to fully understand pathology-related multimodal data, primarily due to the imbalanced data in the pathological domain within their training sets, with pathological images being relatively insufficient compared to the more plentiful textual data. Therefore, we follow LLaVA (Liu et al., 2023) and collect a huge amount of image-caption pairs to align the VLM’s image representation space with that of pathological text. The pretraining data consists of over 700 thousand image-caption pairs from public datasets, websites¹², and from our private annotations on whole-slide images (WSIs) from The Cancer Genome Atlas (TCGA)³.

Furthermore, inspired by contrastive learning in visual training (Khosla et al., 2020; Tian et al., 2020), we conduct pretraining on similarity assessment tasks to enhance the VLM’s capability of obtaining information exclusively from visual input. Specifically, we construct a 200K-sample dataset including two tasks. The first task requires the model to determine whether paired images are semantically equivalent. The second task challenges the model to identify which of the two reference images matches the query image better.

Instruction Following Data The instruction following data covers a diverse range of tasks for downstream applications in computational pathology, including slide-level diagnosis, patch-level subtyping, multiple choice and conversation. These

tasks are carefully selected to address different scenarios of pathological diagnosis. The slide-level diagnosis task involves the recognition and interpretation of global pathological patterns across WSIs, requiring a comprehensive analysis. The patch-level subtyping task focuses on the fine-grained morphological characteristics within regions of interest (RoIs), which emphasizes identifying local features. In addition, the multiple-choice task presents pathological QA and enhances decision-making. Furthermore, the conversation task simulates real-world communication to foster the model’s capability for interaction. Diverse data sources are used to curate the instruction following dataset, which spans public classification datasets, pathological case reports, and our private annotation of TCGA.

3.2 Enriching Context with Multimodal Auxiliary Information

In traditional pathology, pathologists need to integrate slides with patients’ clinical history to make more precise diagnoses. In computational pathology, however, VLMs require even more. Due to the imbalanced pathological data of text over images in VLMs’ domain, terminology explanations are more readily interpretable by VLMs and facilitate the understanding of features from pathological images. Moreover, specific details in pathological images may be too subtle to be adequately represented through text alone. Hence, besides incorporating textual information, learning from visual cues is also significant for VLMs. As Figure 1 illustrates, we explore three strategies to enrich VLMs’ context with such auxiliary information, including clinical history, terminology explanations and visual in-context examples.

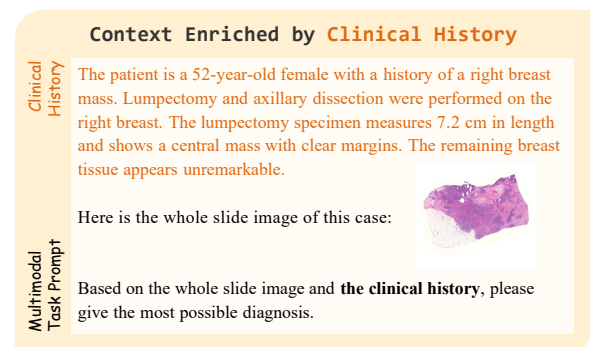


Figure 2: An example prompt with clinical history.

¹<https://hanspopperhepatopathologysociety.org>

²<https://www.webpathology.com>

³<https://portal.gdc.cancer.gov>

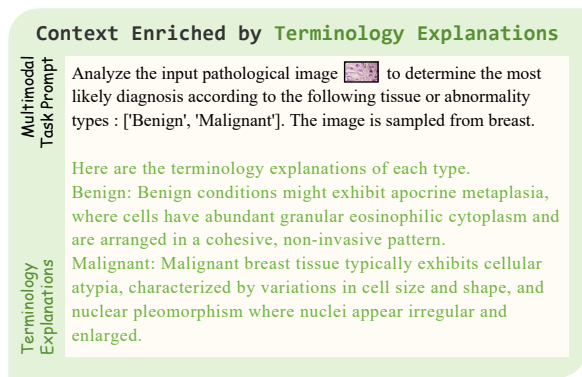


Figure 3: An example prompt with terminology explanations.

3.2.1 Clinical History

Pathological diagnosis relies on multimodal evidence beyond isolated visual patterns. Clinical history plays a crucial role in pathological diagnosis, providing essential context that guides the interpretation of tissue findings. Pathologists can formulate a more accurate diagnosis by considering the patients’ personal information, past medical conditions, symptoms, operations, and biological specimens.

To mimic this phenomenon, we adopt a systematic approach to integrate patients’ clinical history. The clinical history is explicitly included in the task prompt during the inference stage, as illustrated in Figure 2. VLMs subsequently process this textual prompt with the pathological image through their multimodal architecture. In this way, VLMs are guided to simultaneously concentrate their attention on the visual patterns and the relevant clinical history, thereby enabling comprehensive cross-modal analysis.

Before integrating clinical history, data cleaning is necessary, since the raw medical records may contain latent diagnosis leakage. We utilize GPT-4 to analyze and remove explicit or implicit references to the final diagnosis while preserving essential clinical history. Subsequently, the processed clinical history undergoes careful validation by three experienced pathologists, who manually verify the completeness of sensitive information removal and assess the semantic integrity of the remaining content.

3.2.2 Terminology Explanations

A key challenge in applying VLMs for pathological diagnosis is that these models are typically trained on limited pathological images, and struggle when

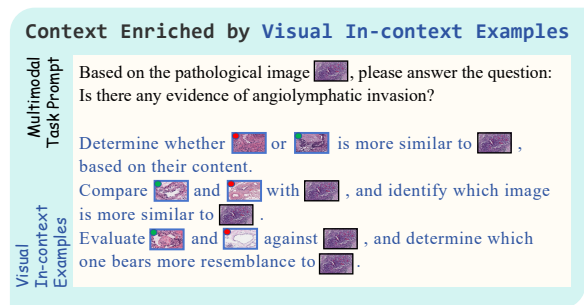


Figure 4: An example prompt with visual in-context examples. The green and red dots in the top-left corner indicate different classes of the reference images.

encountering rare or novel diseases. Pathological texts, however, are relatively plentiful, making it easier for VLMs to interpret them. In this section, we augment VLMs with detailed terminology explanations to address this imbalance and strengthen their domain knowledge. The terminology explanations are the textual descriptions of microscopic pathological images of certain diseases, which focus on morphological characteristics that are more readily interpretable by VLMs. In this way, VLMs are able to comprehend information from pathological images more effectively by establishing connections between visual features and their textual characterizations.

The workflow begins by synthesizing concise terminology explanations with GPT-4, emphasizing cell arrangements, staining patterns, and tissue-level abnormalities. During inference, the terminology explanations of each disease are concatenated to the task prompt and fed into VLMs as a textual input along with the image. Figure 3 presents an example prompt with terminology explanations. By supplementing the image with descriptive context, VLMs benefit from a more comprehensive understanding of each disease category, thus reducing class ambiguity.

3.2.3 Visual In-context Examples

While textual knowledge enhances diagnosis, visual patterns in pathology also serve as critical evidence. For instance, tumor cells present varying appearances across different sites, making it challenging to describe their characteristics solely through text. As a result, it is necessary to learn these characteristics directly from images, since they may contain richer morphological details.

Inspired by in-context learning paradigms, we extend this concept by leveraging images them-

Table 1: Data for pretraining, including data for image-caption alignment and similarity assessment. The data source and size are listed.

	Data Source	Size
Image-caption Alignment	PathGen	300,244
	PathCap	223,169
	Quilt-1M	120,796
	Public Websites	12,325
	TCGA	35,521
Similarity Assessment	TCGA	200,000

selves as contextual guidance for VLMs. In this way, the models are encouraged to allocate greater attention to the visual features of the pathological images when constructing contextual representations. This approach promotes more efficient use of visual patterns for analogical mapping, inductive learning and knowledge generalization. During inference, VLMs evaluate the similarity between the query image and the visual references to determine the class. The classification proceeds by majority voting when multiple reference image pairs are provided. Specifically, the query image is compared with both members of each pair. The label of the selected reference contributes one vote, and the category accumulating the most votes is finally assigned to the query image. Figure 4 demonstrates an example prompt with visual in-context examples. This design aligns with the similarity assessment tasks during fine-tuning, where VLMs have developed capabilities to capture latent visual relationships between image pairs. By prioritizing visual examples, the method emphasizes morphology-driven decision-making.

4 Implementations

We leverage the LLaMA-Factory framework (Zheng et al., 2024) to fine-tune Qwen2-VL-7B-Instruct. We perform the fine-tuning with 3 epochs, which takes about 32 hours on 8 A100 GPUs. The learning rate is set to $1e-4$, and the batch size is set to 8. The warm-up ratio is 0.1.

We list data used in the training phase in Tables 1 and 2. The public datasets of image-caption pairs include PathGen (Sun et al., 2024b), PathCap (Zhang et al., 2020) and Quilt-1M (Ikezogwo et al., 2023). Since PathGen also originates from TCGA, we exclude images that overlap with the testing data. As for slide-level diagnosis, we split

Table 2: Data for instruction following, including data for slide-level diagnosis, patch-level subtyping, multiple choice and conversation. The data source and size are listed.

	Data Source	Size
Slide-level Diagnosis	TCGA Pan-cancer	890
	TCGA-NSCLC	741
	TCGA-RCC	829
Patch-level Subtyping	BreakHis	1,148
	Chaoyang	4,021
	NCT-CRC-HE	100,000
	LC25000	17,500
	PatchGastricADC22	143,497
	PCam	262,144
Multiple Choice	Private Question Bank	7,347
Conversation	TCGA	9,485

TCGA-NSCLC and TCGA-RCC into training and testing sets in a 4:1 ratio, with 741 and 829 training slides, respectively. For patch-level subtyping, we utilize BreakHis (Spanhol et al., 2015), Chaoyang (Zhu et al., 2021), NCT-CRC-HE (Kather et al., 2018), LC25000 (Borkowski et al., 2019), Patch-GastricADC22 (Tsuneki and Kanavati, 2022) and PCam (Veeling et al., 2018), and divide them into training and testing sets according to the official guidance. All data used in this study strictly adhere to the relevant licenses.

In our data preprocessing pipeline, we filter the initial datasets by excluding 80% of the PCam samples and 50% of the Quilt-1M samples, primarily due to concerns regarding image quality and annotation reliability.

During the training and inference phases, all the WSIs are converted to thumbnails to fit the context window limitations of Qwen2-VL.

5 Experiments

Our study evaluates three diagnosis-related tasks that incorporate the three aforementioned types of auxiliary information, respectively. The experimental framework comprises 1) slide-level diagnosis utilizing clinical history, 2) RoI subtyping with terminology explanations, and 3) invasion detection employing image-based in-context learning. Each of the three types of auxiliary information serves a critical role within its corresponding task. For slide-level diagnosis, a comprehensive clinical history helps resolve diagnostic ambiguities inherent

Table 3: Slide-level Diagnosis results(%). w/ H denotes that the model integrates clinical history during inference.

Model	NSCLC		RCC		Pan-cancer	
	Acc	F1	Acc	F1	Acc	F1
Qwen-7B	49.76	43.53	54.84	38.84	10.84	1.56
w/ H	+4.35	+2.45	+1.61	+3.53	+37.50	+43.66
Qwen-72B	48.79	32.00	55.38	38.84	6.40	1.48
w/ H	+14.01	+31.11	+2.15	+5.75	+59.08	+63.01
Ours-7B	73.91	73.88	69.89	67.75	15.71	9.49
w/ H	+1.45	+1.50	+2.15	+1.53	+68.47	+70.74

in large-scale tissue evaluation. RoI subtyping focuses on detailed morphological patterns, where disease-specific terminology explanations provide more targeted guidance than general patient data. In invasion detection, lymphovascular invasion patterns vary across tissue sites, and visual examples enable comparative pattern recognition critical for identifying boundaries.

We implement these experiments using three models, including the baseline Qwen2-VL-7B-Instruct (*Qwen-7B*), Qwen2-VL-72B-Instruct (*Qwen-72B*), and our fine-tuned Qwen2-VL-7B-Instruct (*Ours-7B*).

5.1 Slide-level Diagnosis

Datasets To investigate the impact of clinical history on slide-level diagnosis, we follow Lu et al. (2024a) and conduct evaluations on two classical slide-level subtyping datasets, TCGA-NSCLC (*NSCLC*) and TCGA-RCC (*RCC*), with the test sets containing 207 and 186 WSIs, respectively. An additional class-balanced dataset comprising 904 WSIs of 32 cancer types is constructed from TCGA to comprehensively evaluate the performance of pan-cancer diagnosis (*Pan-cancer*). The test datasets are ensured to be excluded from the training data.

Evaluation Metrics We employ two complementary evaluation metrics, Accuracy (Acc) and F1 score, to assess model performance. Acc provides an intuitive measure of overall diagnostic correctness. The F1 score is simultaneously adopted to measure the performance comprehensively by considering both precision and recall.

Results In Table 3, we present the effect of integrating clinical history. It can be found that systematically integrating such auxiliary information

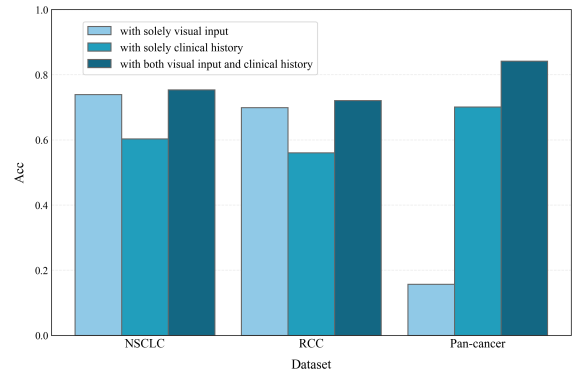


Figure 5: The performances of our fine-tuned model in slide-level diagnosis with solely visual input, with solely clinical history, and with both visual input and clinical history.

leads to notable performance improvements across all datasets. The most substantial improvement is observed on the Pan-cancer dataset, where our fine-tuned model achieves a 68.47% improvement in accuracy. This significant improvement can be attributed mainly to the inclusion of the anatomical site information in the clinical history, which enables the scope reduction of differential diagnosis. Conversely, slide-level subtyping tasks show relatively modest gains, with accuracy improvements of 1.45% on the NSCLC dataset and 2.15% on the RCC dataset. This is mainly because the subtle distinctions in tumor morphology primarily depend more on WSI microstructural pattern recognition than on background clues. Additionally, the Acc and F1 score on Pan-cancer are higher than those on NSCLC and RCC, which also confirms the inherent complexity of tumor-subtyping.

To further study the roles of different modalities, we evaluate our fine-tuned model’s performance with visual input, clinical history, and both infor-

Table 4: RoI Subtyping results (%). w/ E denotes that the model integrates terminology explanations during inference.

Model	BreakHis		Chaoyang		NCT-CRC-HE		LC25000	
	bAcc	wF1	bAcc	wF1	bAcc	wF1	bAcc	wF1
Qwen-7B	50.13	52.08	35.24	34.21	9.51	8.76	20.00	6.85
w/ E	+0.78	+3.02	+3.31	+3.03	+2.88	+6.92	+0.26	+0.76
Qwen-72B	48.56	51.50	29.90	29.11	21.92	14.85	18.26	17.22
w/ E	+1.44	+1.68	+5.96	+4.69	+17.39	+16.99	+1.82	+1.28
Ours-7B	84.00	83.13	70.95	77.86	89.29	90.29	98.98	98.97
w/ E	+5.86	+7.36	-0.58	+0.22	+0.95	+1.16	+0.02	+0.01

mation. From the results shown in Figure 5, we can see that multimodal integration of pathological images with clinical history consistently outperforms unimodal approaches, establishing the necessity of combining visual features with textual context.

5.2 RoI Subtyping

Datasets For the task of RoI subtyping, we use the test dataset of BreakHis, Chaoyang, NCT-CRC-HE, and LC25000, which were excluded from the fine-tuning stage.

Evaluation Metrics To address class imbalance, we apply balanced accuracy (bAcc) and weighted F1 (wF1) score as the evaluation metrics.

Results Table 4 demonstrates a notable improvement in performance tied to terminology explanation supplementation. The most pronounced improvements are observed in the Chaoyang and NCT-CRC-HE datasets for the baseline Qwen-VL models. This enhancement primarily stems from the insufficient prior domain knowledge of Qwen-VL in these pathological subtypes. In contrast, our fine-tuned model achieves substantially higher baseline performance, which confirms that domain adaptation contributes to robust feature comprehension. Nevertheless, terminology explanations still yield non-trivial additional gains on two of the four datasets, highlighting their effectiveness even in optimized models.

In summary, these findings indicate that integrating terminology explanations is an efficient way to inject prior knowledge. Models can recognize discriminative features more effectively by supplying terminology explanations that outline each disease category with morphological descriptions. These added cues compensate for the original knowledge

gap and allow the model to generalize more reliably across unseen cases.

5.3 Invasion Detection

Datasets The development of invasion detection remains challenging due to the lack of publicly available datasets. We establish a private dataset comprising 596 samples from the Pan-cancer dataset to address the issue. The dataset contains 376 negative cases (absence of invasion) and 220 positive cases (confirmed invasion presence), with all samples undergoing rigorous annotation by three experienced pathologists.

Table 5: Invasion Detection results (%). w/ n -shots denotes that the model integrates n pairs of visual in-context examples during inference. The best performance in each column is bold.

Model	Invasion Detection		
	Acc	F1	Recall
Qwen-7B	62.08	9.56	5.91
Qwen-72B	59.39	58.12	76.36
Ours-7B	77.18	57.50	41.82
Ours-7B w/ 1-shot	81.86	71.23	70.91
Ours-7B w/ 5-shot	87.25	81.82	77.73
Ours-7B w/ 10-shot	89.09	84.19	78.64

Evaluation Metrics In addition to the Acc and F1 score, we further incorporate Recall as a critical evaluation metric for the invasion detection task. Recall directly quantifies models' ability to detect true invasion cases, which constitutes the primary objective in actual applications. We omit the Precision metrics due to space limitations and the

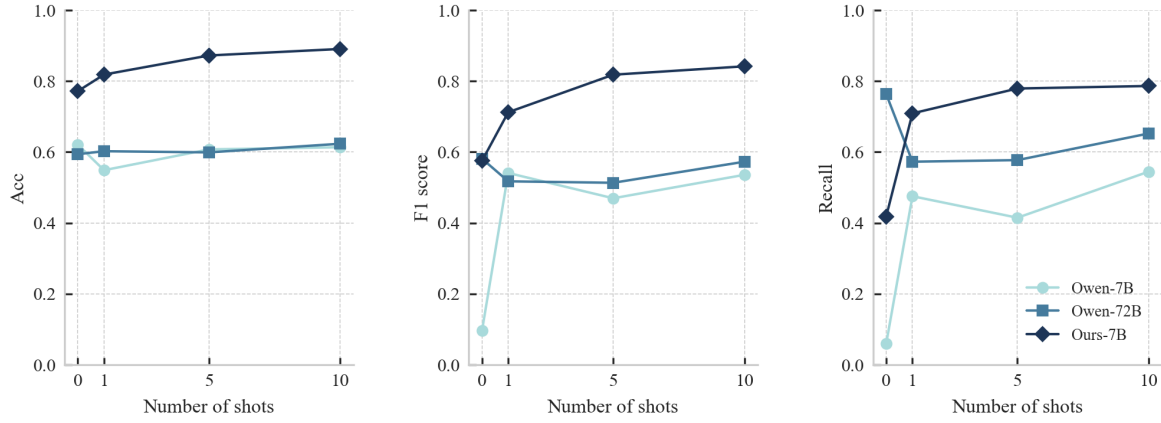


Figure 6: Comparison of model performance (Acc, F1 score, and Recall) in invasion detection regarding different numbers of shots.

prioritization of recall over precision in this task.

Results Table 5 demonstrates our experiment results in invasion detection. Notably, parameter scaling in the baseline Qwen-VL architecture fails to yield consistent performance improvements. This phenomenon also highlights the necessity of domain-specific fine-tuning, which significantly enhances the capabilities of the model (+47.94% F1 score versus the baseline model).

Furthermore, our fine-tuned model exhibits adaptive behavior when processing visual in-context examples. Under zero-shot conditions, the model achieves 77.18% accuracy but demonstrates limited recall of 41.82%, which suggests a conservative classification strategy where predictions are made only for high-confidence cases, resulting in missed positive samples. Introducing one-shot visual examples triggers a strategic adaptation. Recall increases substantially to 70.91%, indicating that visual contextual guidance enables more active identification of positive cases. With extended context (5/10-shot configurations), the performance further improves, **achieving 87.25% and 89.09% accuracy while maintaining high recall rates (77.73% and 78.64%)**. This progression suggests that the increased visual contextual information enhances task comprehension and strengthens the model’s ability to generalize features from positive samples. To illustrate this point more concretely, a case study is included in Appendix A.3.

Besides, Figure 6 demonstrates that when baseline Qwen2-VL models are provided with visual in-context examples, their gains are far more muted than those observed for our fine-tuned model (even ineffective). For the Qwen2-VL-7B-Instruct model,

applying image-based in-context learning significantly improves the recall and F1 score compared to zero-shot scenarios. However, this performance gain remains stagnant as the number of shots increases, indicating the model’s limited capacity in learning visual features from pathological images through incremental examples. Similarly, the Qwen2-VL-72B-Instruct model exhibits negligible improvement or even slight degradation in recall and F1 score with few-shot demonstrations, further confirming the observed limitations in visual feature summarization. Importantly, both the baseline and our fine-tuned Qwen-VL model display gain saturation when exceeding a threshold of shot numbers.

6 Conclusion

This study establishes a paradigm for advancing computational pathology with multimodal auxiliary information. By systematically incorporating clinical history, terminology explanations and visual in-context examples, we propose a context-enriched framework that addresses the inherent limitations of conventional VLMs in pathological diagnosis. Our experiments across slide-level diagnosis, RoI subtyping and invasion detection validate the effectiveness of this approach. The findings indicate that bridging the expertise gap between human experts and artificial systems is not a purely architectural problem but a contextual one. By providing VLMs with auxiliary information, we can move a step closer to trustworthy computational pathology. Future work will address automatic retrieval of relevant auxiliary information and validation within clinical practice.

Limitations

There are two main limitations in our work. First, we conduct fine-tuning on a general-purpose VLM for developing pathology-specific applications, rather than pursuing an alternative approach following LLaVA. The latter approach aligns the visual and textual embedding space of domain-optimized image encoders and large language models, which leverages the powerful feature extraction capabilities of pretrained transformers for pathological diagnosis. We do not try this approach because we do not have enough high-quality pathological image-caption pairs and QA datasets. As a reference, PathChat utilizes 1.18 million image-caption pairs and over 450 thousand instructions, and they don't release these training data publicly. While our general-purpose VLM-based approach benefits from pretrained cross-modal representations, the model has inherent limitations in the interpretation of pathological images.

Secondly, we conduct the evaluation exclusively on standardized benchmark datasets without real-world clinical validation. While the benchmark datasets could provide a more quantitative comparison, they may not fully capture the complexity, variability, and noise in real-world clinical environments.

References

- Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. 2019. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, and 1 others. 2024a. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862.
- Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. 2024b. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. *arXiv preprint arXiv:2410.11761*.

- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, and 1 others. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.
- Dyke Ferber, Georg Wölflein, Isabella C Wiest, Marta Ligeró, Srividhya Sainath, Narmin Ghaffari Laleh, Omar SM El Nahhas, Gustav Müller-Franzes, Dirk Jäger, Daniel Truhn, and 1 others. 2024. In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, 15(1):10104.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. 2023. A visual-language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316.
- Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2023. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017.
- Tianlei Jin, Fangtai Guo, Qiwei Meng, Shiqiang Zhu, Xiangming Xi, Wen Wang, Zonghao Mu, and Wei Song. 2023. Fast contextual scene graph generation with unbiased context augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6302–6311.
- Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 2018. [100,000 histological images of human colorectal cancer and healthy tissue](#).
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.
- Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. [Can language models learn from explanations in context?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. 2023. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in*

650	<i>neural information processing systems</i> , 36:34892–	Yuxuan Sun, Yixuan Si, Chenglu Zhu, Xuan Gong,	703
651	34916.	Kai Zhang, Pingyi Chen, Ye Zhang, Zhongyi Shui,	704
652	Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan,	Tao Lin, and Lin Yang. 2024a. Cpath-omni: A	705
653	Lawrence Carin, and Weizhu Chen. 2021. What	unified multimodal foundation model for patch and	706
654	makes good in-context examples for gpt-3? <i>arXiv</i>	whole slide image analysis in computational pathol-	707
655	<i>preprint arXiv:2101.06804</i> .	ogy. <i>arXiv preprint arXiv:2412.12077</i> .	708
656	Shih-Wen Liu, Hsuan-Yu Fan, Wei-Ta Chu, Fu-En Yang,	Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu,	709
657	and Yu-Chiang Frank Wang. 2025. Histopathology	Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng	710
658	image report generation by vision language model	Lyu, Tao Lin, and Lin Yang. 2024b. Pathgen-1.6	711
659	with multimodal in-context learning. In <i>Medical</i>	m: 1.6 million pathology image-text pairs generation	712
660	<i>Imaging with Deep Learning</i> .	through multi-agent collaboration. <i>arXiv preprint</i>	713
661	Ming Y Lu, Bowen Chen, Drew FK Williamson,	<i>arXiv:2407.00203</i> .	714
662	Richard J Chen, Ivy Liang, Tong Ding, Guillaume	Shounak Sural, Nishad Sahu, and Ragunathan Raj Ra-	715
663	Jaume, Igor Odintsov, Long Phi Le, Georg Gerber,	jkumar. 2024. Contextualfusion: Context-based	716
664	and 1 others. 2024a. A visual-language founda-	multi-sensor fusion for 3d object detection in ad-	717
665	tion model for computational pathology. <i>Nature</i>	verse operating conditions. In <i>2024 IEEE intelligent</i>	718
666	<i>Medicine</i> , 30(3):863–874.	<i>vehicles symposium (IV)</i> , pages 1534–1541. IEEE.	719
667	Ming Y Lu, Bowen Chen, Drew FK Williamson,	Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Feng-	720
668	Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji	tao Zhou, Yi Zhang, and Bo Liu. 2023. Multiple	721
669	Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Pa-	instance learning framework with masked hard in-	722
670	tel, and 1 others. 2024b. A multimodal generative ai	stance mining for whole slide image classification.	723
671	copilot for human pathology. <i>Nature</i> , 634(8033):466–	In <i>Proceedings of the IEEE/CVF International Con-</i>	724
672	473.	<i>ference on Computer Vision</i> , pages 4078–4087.	725
673	Ming Y Lu, Drew FK Williamson, Tiffany Y Chen,	Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan,	726
674	Richard J Chen, Matteo Barbieri, and Faisal Mah-	Cordelia Schmid, and Phillip Isola. 2020. What	727
675	mood. 2021a. Data-efficient and weakly supervised	makes for good views for contrastive learning? <i>Ad-</i>	728
676	computational pathology on whole-slide images. <i>Nature</i>	<i>advances in neural information processing systems</i> ,	729
677	<i>biomedical engineering</i> , 5(6):555–570.	33:6827–6839.	730
678	Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel,	Masayuki Tsuneki and Fahdi Kanavati. 2022. Inference	731
679	and Pontus Stenetorp. 2021b. Fantastically ordered	of captions from histopathological patches. In <i>Inter-</i>	732
680	prompts and where to find them: Overcom-	<i>national Conference on Medical Imaging with Deep</i>	733
681	ing few-shot prompt order sensitivity. <i>arXiv preprint</i>	<i>Learning</i> , pages 1235–1250. PMLR.	734
682	<i>arXiv:2104.08786</i> .	Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco	735
683	Subash Neupane, Shaswata Mitra, Sudip Mittal, Manas	Cohen, and Max Welling. 2018. Rotation equivariant	736
684	Gaur, Noorbakhsh Amiri Golilarz, Shahram Rahimi,	cnns for digital pathology. In <i>Medical image com-</i>	737
685	and Amin Amirlatifi. 2024. Medinsight: A multi-	<i>puting and computer assisted intervention–mICCAI</i>	738
686	source context augmentation framework for gener-	2018: 21st international conference, granada, Spain,	739
687	ating patient-centric medical responses using large	September 16-20, 2018, <i>proceedings, part II 11</i> ,	740
688	language models. <i>ACM Transactions on Computing</i>	pages 210–218. Springer.	741
689	<i>for Healthcare</i> .	Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhi-	742
690	Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan,	hao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin	743
691	Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan	Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei	744
692	Larson, Yuanzhi Li, Weishung Liu, and 1 others.	Du, Xuancheng Ren, Rui Men, Dayiheng Liu,	745
693	2023. Can generalist foundation models outcom-	Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a.	746
694	pete special-purpose tuning? case study in medicine.	Qwen2-vl: Enhancing vision-language model’s per-	747
695	<i>arXiv preprint arXiv:2311.16452</i> .	ception of the world at any resolution . <i>Preprint</i> ,	748
696	Gabriel Recchia. 2021. Teaching autoregressive lan-	<i>arXiv:2409.12191</i> .	749
697	guage models complex tasks by demonstration.	Yubo Wang, Xueguang Ma, and Wenhui Chen. 2024b.	750
698	<i>arXiv preprint arXiv:2109.02102</i> .	Augmenting black-box llms with medical textbooks	751
699	Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean,	for biomedical question answering. In <i>Findings</i>	752
700	and Laurent Heutte. 2015. A dataset for breast cancer	<i>of the Association for Computational Linguistics:</i>	753
701	histopathological image classification. <i>Ieee transac-</i>	<i>EMNLP 2024</i> , pages 1754–1770.	754
702	<i>tions on biomedical engineering</i> , 63(7):1455–1462.	Patrick H Winston. 1980. Learning and reasoning by	755
		analogy. <i>Communications of the ACM</i> , 23(12):689–	756
		703.	757

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. [Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, and 1 others. 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, and 1 others. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):A10a2300068.

Renyu Zhang, Christopher Weber, Robert Grossman, and Aly A Khan. 2020. Evaluating and interpreting caption prediction for histopathology images. In *Machine Learning for Healthcare Conference*, pages 418–435. PMLR.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Chuang Zhu, Wenkai Chen, Ting Peng, Ying Wang, and Mulan Jin. 2021. Hard sample aware noise robust learning for histopathology image classification. *IEEE transactions on medical imaging*, 41(4):881–894.

A Appendix

A.1 Terminology Explanations in RoI Subtyping

Table 6 lists the terminology explanations used in RoI Subtyping.

A.2 Data Samples of the Fine-tuning Data

Table 7 lists the data samples of the fine-tuning data.

A.3 Case Study

Slide-level Diagnosis Figure 7 presents our fine-tuned model’s responses to a case from the Pancancer dataset. As shown, the model mistakenly interprets the slide as thyroid carcinoma at the beginning. By integrating clinical history, the model can form a more accurate diagnosis.












RoI Subtyping Figure 8 shows our fine-tuned model’s responses to a case from the BreakHis dataset. With the incorporation of terminology explanations, the model corrects its answer and successfully identifies the malignant breast cancer.

Invasion Detection Figure 9 demonstrates the effect of visual in-context examples in invasion detection. Although our fine-tuned model is provided with 1-shot visual in-context examples, it still fails to detect the invasion, which is consistent with the limited recall observed in our experimental results. By integrating 5-shot visual in-context examples, the model is able to identify the invasion. This case highlights the effectiveness of visual in-context examples for enhancing task comprehension.

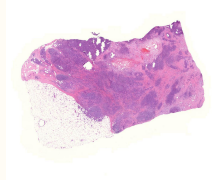
Table 6: The terminology explanations used in RoI subtyping.

Dataset	Terminology Explanations
BreakHis	Benign: Benign conditions might exhibit apocrine metaplasia, where cells have abundant granular eosinophilic cytoplasm and are arranged in a cohesive, non-invasive pattern.
	Malignant: Malignant breast tissue typically exhibits cellular atypia, characterized by variations in cell size and shape, and nuclear pleomorphism where nuclei appear irregular and enlarged.
Chaoyang	Normal: Normal colon tissue shows a balance of epithelial cells making up the crypts and absorptive cells, with an orderly appearance devoid of dysplasia or atypical architecture. The muscularis mucosae are typically seen at the base, underlying the crypts.
	Adenocarcinoma: Colonic adenocarcinoma presents with crowded, back-to-back glands that often show significant nuclear atypia, loss of mucin production, and occasionally areas of necrosis.
	Adenoma: Adenomas in the colon typically feature glandular structures that are closely packed together. The cells within these glandular formations often exhibit nuclear atypia, with elongated, hyperchromatic nuclei that are stratified.
	Serrated: In the context of the colon, serrated tissue abnormalities present a jagged or serrated pattern along the glands. These formations can be further divided into hyperplastic polyps, sessile serrated lesions, and traditional serrated adenomas based on their microscopic appearance.
NCT-CRC-HE	Adipose (ADI): Adipose tissue in colorectal cancer and normal tissue typically appears as clusters of empty-looking cells with clear cytoplasm and thin, peripheral nuclei, due to the dissolution of lipid content during processing.
	Background (BACK): In histological images from colorectal cancer (CRC) and normal tissue, the background can include areas of smooth muscle, normal epithelial cells, and immune cell infiltrates.
	Debris (DEB): DEB in CRC and normal tissue often appears as amorphous, eosinophilic material scattered within the tissue sections. It can include remnants of necrotic cells, tissue fragments, and extracellular material.
	Lymphocytes (LYM): In the context of colorectal cancer and normal tissue, LYM can often be seen infiltrating the tumor microenvironment. Visually, they are characterized by their small size and round, dark nuclei with a scant cytoplasm.
	Mucus (MUC): In normal colorectal tissue, MUC is typically secreted by goblet cells within the epithelial lining of the colon and rectum. These cells have a characteristic appearance with a distended, mucin-filled cytoplasm and a small, compressed nucleus at the base of the cell.
	Smooth Muscle (MUS): In both CRC and normal tissue, MUS appears as elongated, spindle-shaped cells with centrally located nuclei. The cells are often arranged in parallel bundles or sheets and may show a characteristic wavy pattern.
	Normal Colon Mucosa (NORM): The lamina propria in normal colon mucosa is sparse and contains a relatively low density of inflammatory cells. This tissue section exhibits a smooth regular surface with intact epithelial cells, featuring a consistent arrangement without significant distortion.
	Cancer-associated Stroma (STR): Stromal areas in CRC may include inflammatory cell infiltrates, which can be observed as clusters of immune cells interspersed within the fibrous tissue.
LC25000	Colorectal Adenocarcinoma Epithelium (TUM): TUM often exhibits aberrant glandular architecture with prominent nucleoli and hyperchromatic cells. There is usually a disruption in the normal glandular arrangement, with glands appearing haphazardly distributed.
	Lung adenocarcinomas: Lung adenocarcinomas typically present as irregularly shaped glands and acinar structures, often exhibiting mucin production. The cells can be columnar or cuboidal and usually have prominent nucleoli.
	Benign colonic tissues: Benign colonic tissues are characterized by the presence of well-organized, tubular glands lined by uniform columnar epithelial cells with basally located nuclei.
	Colon adenocarcinomas: In colon adenocarcinomas, the gland formations are frequently distorted and crowded, with back-to-back glands and a cribriform pattern. The cells within these glands often display atypical nuclei and mitotic figures.
	Lung squamous cell carcinomas: In lung squamous cell carcinomas, the tissue typically shows irregular, infiltrative growth patterns with stratified squamous cells that may exhibit keratinization.
LC25000	Benign lung tissues: The stroma in benign lung tissues appears normal, with no evidence of fibrosis, desmoplasia, or inflammatory infiltrate. The interstitial spaces are clear, and the overall tissue architecture maintains its standard form without masses or lesions.

Table 7: Data samples of the fine-tuning data.

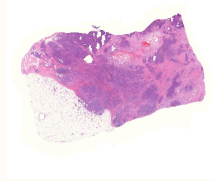
	Data Samples
Image-caption Alignment	<p>Question: Tell me what is shown in this pathological image . Summarize your findings in simple sentences or key words.</p> <p>Answer: Histopathological examination (hematoxylin and eosin) of the sino-atrial node shows interstitial fibrosis and eosinophilic inflammatory infiltrate at 40x magnification.</p>
Similarity Assessment	<p>Question: Compare the given Image A  and Image B , and decide whether these two represent the same region.</p> <p>Answer: No.</p> <p>Question: Determine whether Image A  or Image B  is more similar to Image C , based on their content.</p> <p>Answer: Image B.</p>
Slide-level Diagnosis	<p>Question: Based on the whole slide image, please give the most possible diagnosis. </p> <p>Answer: Lung adenocarcinoma.</p>
Patch-level Subtyping	<p>Question: Analyze the input pathological image  to determine the most likely diagnosis according to the following tissue or abnormality types: ['Benign', 'Malignant'].</p> <p>Answer: Malignant.</p>
Multiple Choice	<p>Question: What is your diagnosis based on the image ?</p> <p>A. Tubular adenoma with high grade dysplasia; B. Tubular adenoma with invasive adenocarcinoma; C. Tubular adenoma with invasive adenocarcinoma with associated mucin production; D. Tubular adenoma with pseudoinvasion.</p> <p>Answer: D. Tubular adenoma with pseudoinvasion.</p>
Conversation	<p>Question1: Based on the whole slide image, please give the most possible diagnosis. </p> <p>Answer1: Papillary thyroid carcinoma.</p> <p>Question2: This is a crop  from the slide. Please answer the question based on previous analysis and the crop. Is there evidence of papillary variants in the image?</p> <p>Answer2: Yes, there is clear evidence.</p>

Original Diagnosis

<div style="writing-mode: vertical-rl; transform: rotate(180deg);">Multimodal Task Prompt</div>	<p>Here is the whole slide image of this case:</p> <div style="text-align: center; margin: 10px 0;">  </div> <p>Based on the whole slide image, please give the most possible diagnosis.</p>
<div style="writing-mode: vertical-rl; transform: rotate(180deg);">Model's Response</div>	<p>Thyroid carcinoma.</p>

(a) The original diagnosis.

Context Enriched by Clinical History


<div style="writing-mode: vertical-rl; transform: rotate(180deg);">Clinical History</div>	<p>The patient is a 52-year-old female with a history of a right breast mass. Lumpectomy and axillary dissection were performed on the right breast. The lumpectomy specimen measures 7.2 cm in length and shows a central mass with clear margins. The remaining breast tissue appears unremarkable.</p>
<div style="writing-mode: vertical-rl; transform: rotate(180deg);">Multimodal Task Prompt</div>	<p>Here is the whole slide image of this case:</p> <div style="text-align: center; margin: 10px 0;">  </div> <p>Based on the whole slide image and the clinical history, please give the most possible diagnosis.</p>
<div style="writing-mode: vertical-rl; transform: rotate(180deg);">Model's Response</div>	<p>Breast invasive carcinoma.</p>

(b) The enhanced diagnosis by enriching context with clinical history.

Figure 7: Case illustration of enriching context in slide-level diagnosis. The ground truth is *breast invasive carcinoma*.

Original Diagnosis

Multimodal Task Prompt

Analyze the input pathological image  to determine the most likely diagnosis according to the following tissue or abnormality types : ['Benign', 'Malignant']. The image is sampled from breast.


Model's Response

Benign.

(a) The original diagnosis.

Context Enriched by Terminology Explanations

Multimodal Task Prompt

Analyze the input pathological image  to determine the most likely diagnosis according to the following tissue or abnormality types : ['Benign', 'Malignant']. The image is sampled from breast.

Terminology Explanations

Here are the terminology explanations of each type.

Benign: Benign conditions might exhibit apocrine metaplasia, where cells have abundant granular eosinophilic cytoplasm and are arranged in a cohesive, non-invasive pattern.

Malignant: Malignant breast tissue typically exhibits cellular atypia, characterized by variations in cell size and shape, and nuclear pleomorphism where nuclei appear irregular and enlarged.

Model's Response


Malignant.

(b) The enhanced diagnosis by enriching context with terminology explanations.

Figure 8: Case illustration of enriching context in RoI subtyping. The ground truth is *malignant*.

Original Diagnosis

Multimodal Task Prompt

Based on the pathological image , please answer the question: Is there any evidence of lymphovascular invasion?




Model's Response

No, there is no evidence of lymphovascular invasion.

(a) The original diagnosis.

Context Enriched by Visual In-context Examples (1-shot)

Visual In-context Examples

Determine whether Image A  or Image B  is more similar to Image C , based on their content.




Model's Response




Image B. (●)




(b) The wrong diagnosis despite the integration of 1-shot visual in-context examples.




Context Enriched by Visual In-context Examples (5-shot)




Visual In-context Examples

Determine whether Image A  or Image B  is more similar to Image C , based on their content.

Compare Image A  and Image B  with Image C , and identify which image is more similar to Image C.

Evaluate Image A  and Image B  against Image C , and determine which one bears more resemblance to Image C.

Determine whether Image A  or Image B  is more similar to Image C , based on their content.

Compare Image A  and Image B  with Image C , and identify which image is more similar to Image C.

Model's Response

Image B. (●)

Image B. (●)

Image A. (●)

Image A. (●)

Image A. (●)

(● ● ● ● ● The final diagnosis is existence of invasion.)

(c) The enhanced diagnosis by enriching context with 5-shot visual in-context examples.

Figure 9: Case illustration of enriching context in invasion detection. The ground truth is that *there is evidence of invasion*. The green dot indicates the existence of an invasion, while the red dot indicates the absence. The content within the parentheses is provided solely for clarity and does not represent the model's output.