

CPath-Omni: A Unified Multimodal Foundation Model for Patch and Whole Slide Image Analysis in Computational Pathology

Supplementary Material

A. Additional Experiments and Details

A.1. Ablations of Vision and Text Components in CPath-CLIP

We further explore the influence of different vision and text components on CPath-CLIP’s performance in zero-shot tasks to explore its semantic alignment capabilities and understand the role of each element. Our experiments include evaluating CLIP-L alone, Virchow2 alone, and a combination of both, as well as fixing the vision encoder and comparing text encoders by substituting CLIP-L with Qwen2-1.5B. As shown in Tab. B.1, when using Virchow2 as the fixed vision backbone and replacing the CLIP-L text encoder with Qwen2-1.5B, we observe a 0.9% overall performance improvement. Conversely, fixing the text encoder as Qwen2-1.5B and replacing CLIP-L with Virchow2 results in a significant 13.7% performance increase. This suggests that the primary boost is attributed to Virchow2’s pathology-specific pretraining on 3.1 million whole-slide images, highlighting that a more advanced pathology encoder greatly enhances semantic alignment capabilities. Furthermore, combining CLIP-L with Virchow2 provides an additional 0.3% performance boost. While this gain is modest compared to the standalone Virchow2 encoder, we retain it to enrich semantic features for future integration into LLM.

A.2. CPath-Omni Performance in Patch and WSI Classification Tasks

Tab. B.2, Tab. B.3, and Tab. B.4 present detailed metrics for patch-level and WSI-level classification corresponding to the radar plot visualization shown in Fig. 4 of the main paper. We also compare state-of-the-art pathology LMMs, Quilt-LLaVA, and PathGen-LLaVA. Notably, these models cannot directly perform WSI classification. To adapt them for this task, we employ the same method used with GPT-4o in the main paper: generating captions for individual patches and merging them into a WSI-level report, which is then used for classification based on predefined questions. Further details are available in Section 5.3 of the main paper.

Our findings show that CPath-Omni significantly outperforms previous models in both patch-level and WSI-level classification tasks. Notably, in out-of-distribution or zero-shot patch classification datasets (Tab. B.3), none of the models were trained on these datasets, making it a relatively fairer comparison. In this context, CPath-Omni signifi-

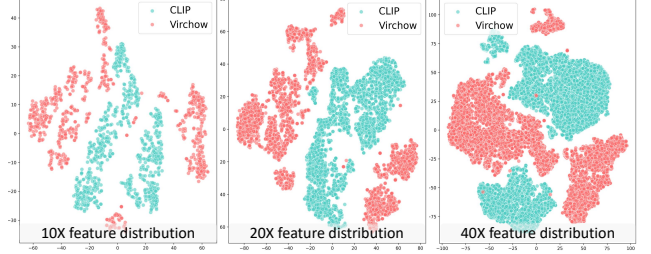


Figure A.1. Visualization of CLIP and Virchow2 feature distributions within CPath-CLIP for a whole slide image.

cantly surpasses GPT-4o and Gemini-1.5-Pro, the strongest general models, as well as pathology-specific LMMs such as PathGen-LLaVA and Quilt-LLaVA. Additionally, CPath-Omni achieves performance comparable to task-specific fine-tuned models, underscoring its strength in task unification and exceptional overall performance.

A.3. Feature Visualization of the Two Vision Encoders in CPath-CLIP

As CPath-CLIP contains both a CLIP vision encoder and a Virchow2 vision encoder, the two encoders correspond to different dimensions of the same Qwen2-1.5B text embedding (Virchow2 maps to the first 2560 features, while CLIP maps to features 2560-3328), each capturing different characteristics. To demonstrate that these vision encoders extract distinct visual features, we visualized the CLIP and Virchow2 features extracted from a whole slide image using PCA and t-SNE, as shown in Fig. A.1. The visualization confirms their distinct feature distributions.

A.4. Experiment Details of Patch-level Linear Probing and Task-Specific Model Fine-Tuning

Linear Probing: The linear probing experiment evaluates the representational power of a pre-trained model by adding a linear layer to its output. This linear layer maps the model’s output vector to the number of classes, enabling classification. The experiment uses a batch size of 32 and runs for 20 epochs. The optimizer is AdamW with a learning rate of 1×10^{-2} . To ensure robustness and reproducibility, we employ 10 different seeds. The procedure involves randomly selecting N samples (2, 8, 16, 32, 64, 128) from each class to form the training set. If an official test set is unavailable or lacks labels, the remainder of the dataset serves as the test set. Throughout the 20 epochs, we select the best-performing model based on its accuracy on the test

set, providing insights into the effectiveness of the added linear layer in classifying unseen data.

Task-specific model fine-tuning: For task-specific model fine-tuning, we build on the linear probing setup by unfreezing the Virchow2 backbone and performing full-parameter fine-tuning on the model using the entire training set.

A.5. Details for WSI preprocessing

For WSI preprocessing, we utilize slightly modified CLAM to identify and segment regions by setting appropriate thresholds. Within each WSI, we extract 2048×2048 non-overlapping patches at a magnification of $40\times$ from the identified regions. Patches are retained if more than 10% of their area contains valid tissue regions. Additionally, each 2048×2048 patch is further subdivided into one 2048×2048 patch, four 1024×1024 patches, and sixteen 512×512 patches for subsequent feature extraction.

A.6. Details for WSI Task-Specific Fine-Tuning

The models are trained for 20 epochs without a learning rate schedule, using a fixed learning rate of 1×10^{-5} . The training process utilizes the Adam optimizer without weight decay, and the batch size is consistently set to 1.

Model Architecture. The MIL framework commonly used for WSI classification includes three learnable components: (1) A fully-connected layer to reduce the dimensionality of features to 256. (2) An attention network to aggregate and transform the instance features. (3) A final fully-connected layer for making predictions. We experiment with ABMIL and DSMIL. Both models share the same fully connected layers for reducing feature dimensionality and making predictions. For the attention network, ABMIL uses the gated attention mechanism, while DSMIL introduces a dual-stream architecture.

For a fair comparison, the input patch features of these two models are kept consistent with those of CPath-Omni.

A.7. Training Hyperparameters for CPath-CLIP and CPath-Omni

We detail the training parameters of CPath-CLIP in Tab. B.6. The hyperparameters for the four training stages of CPath-Omni are listed in Tab. B.7, Tab. B.8, Tab. B.9, and Tab. B.10, respectively. Specifically, stages 1 and 2 focus on patch-based training, stage 3 is dedicated to WSI-based training, and stage 4 involves a mix of patch-based and WSI-based training.

A.8. Hardware

We employ 8 NVIDIA H800-80G GPUs to train the CPath-Omni model, with a total training time of 45 hours across all stages (Stage 1: 8.5h, Stage 2: 25h, Stage 3: 0.5h, Stage 4: 11h). 1 NVIDIA A100-40G GPU for fine-tuning task-

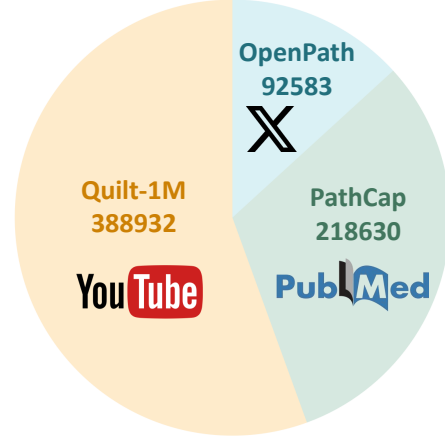


Figure B.2. Proportions of sub-datasets in CPath-PatchCaption and their primary sources.

specific models, and 4 NVIDIA H800-80G GPUs for caption generation using PathGen-LLaVA and Quilt-LLaVA.

B. Additional Details of Collected Datasets

We provide details on the dataset sources and distribution of CPath-PatchCaption in Fig. B.2. The sources, quantities, distributions, and sub-task data allocations for CPath-PatchInstruction and CPath-WSIInstruction are also illustrated in Fig. B.3 and Fig. B.4.

Additionally, the construction of CPath-VQA within the CPath-Instruct dataset follows a systematic approach. First, we collect datasets that already include captions. For datasets lacking captions, such as classification datasets, we use GPT-4o to generate captions by combining classification labels with image data. GPT-4o further generates VQA pairs from these captioning datasets, creating the CPath-VQA.

We also present visualization examples of the novel task of visual referring prompting, alongside tasks specific to whole-slide images, which differ from natural images due to their extremely high resolution, reaching nearly $100,000 \times 100,000$ pixels. Specifically, Fig. B.5 illustrates the annotation interface for visual referring prompting and shows how pathologists annotate this task. Fig. B.6 provides an example of generating visual referring prompting. Fig. B.7 showcases examples of WSI captioning and VQA. In these examples, we present the cleaned report, which, after processing, aligns well with the pathological representations in the whole-slide image. We highlight the correspondences between features in the report and the whole-slide image for better clarity. Based on this cleaned report, we generate both multiple-choice and open-ended examples.

Vision Encoder		Text Encoder	LC-Lung	LC-Colon	CRC100K	SkinCancer	Pcam	BACH	Osteo	WSSSLUAD	SICAPv2	Average
CLIP-L	Virchow2											
✗	✓	CLIP-L	94.9	99.5	77.7	70.1	91.1	74.5	72.7	89.9	67.9	82.0
✗	✓	Qwen2-1.5B	96.8	99.8	76.0	78.1	94.6	72.0	82.0	86.4	60.7	82.9
✓	✗	Qwen2-1.5B	92.1	91.9	67.0	61.7	87.8	47.5	52.9	79.0	42.9	69.2
✓	✓	Qwen2-1.5B	97.1	100.0	78.0	74.2	95.9	72.3	80.7	87.1	63.1	83.2

Table B.1. Zero-shot classification performance comparison of CPath-CLIP built with different vision and text encoders.

	GPT-4o	Gemini-1.5-pro	Quilt-LLaVA	PathGen-LLaVA	Full-finetune	CPath-Omni (ours)
VALSET_TCGA	39.2	28.0	23.5	28.9	97.0	<u>96.0</u>
Stomach	18.7	33.3	21.6	19.6	83.2	<u>82.6</u>
KIRC	79.8	84.4	38.2	90.8	<u>99.4</u>	99.6
CocaHis	90.0	60.0	43.4	89.7	88.0	90.0
WISEPAIP23	30.0	47.8	14.5	15.9	89.4	<u>88.4</u>
VALSET_WNS	29.5	26.2	17.3	23.3	93.8	<u>91.2</u>
BCNB	52.0	65.8	65.7	65.2	90.4	<u>90.0</u>
VALSET_CHA	30.2	30.6	16.2	25.2	96.6	<u>93.2</u>
CATCH	20.4	36.0	26.7	26.7	<u>79.0</u>	83.2
PAIP21	7.0	37.4	18.9	18.1	92.6	<u>86.2</u>
MIDOG22	47.7	62.1	50.9	48.9	<u>65.8</u>	80.2
KICH	74.0	73.0	29.6	87.1	<u>99.4</u>	100.0
CAMEL	64.2	53.4	56.7	61.0	<u>91.4</u>	92.4
Gleason_CNN	46.3	51.8	38.8	39.0	81.7	81.7
OCELOT	44.2	27.3	20.9	37.2	90.9	<u>84.4</u>
Prostate (Tolkach Y et al.)	62.8	54.4	50.4	54.5	<u>97.6</u>	99.0
Average	46.0	48.2	33.3	45.7	<u>89.8</u>	89.9

Table B.2. Performance comparison of general-purpose, pathology-specific, and task-specific models on ID patch classification tasks.

C. Prompts for GPT-4o

This section presents all the prompts used in our dataset and experimental process, including: (1) the prompt in Fig. C.8, which is used with GPT-4o to enrich and refine existing image captions by adding details; (2) the prompt in Figure Fig. C.9, which is utilized to modify raw WSI reports by removing information that cannot be directly observed in the WSI, such as gross specimen descriptions; (3) the prompts in Fig. C.10 and Fig. C.11, which are applied to generate closed-ended and open-ended VQA pairs based on cleaned WSI reports; (4) the prompt in Fig. C.12, which is designed to prompt GPT-4o to generate captions for image patches within a WSI; (5) the prompt in Fig. C.13, which is used to merge all the captions generated for individual patches in a WSI into a cohesive generated WSI report; (6) the prompts in Fig. C.14 and Fig. C.15, which are employed to guide GPT-4o in answering closed-ended and open-ended WSI VQAs by analyzing whether the answers can be derived from the generated WSI report; and (7) the prompt in Figure C.16, which is used to determine whether the answer to an open-ended WSI VQA is correct by referencing the provided question and answer.

	GPT-4o	Gemini-1.5-pro	Quilt-LLaVA	PathGen-LLaVA	CPath-Omni (ours)
SkinCancer	33.7	30.2	6.7	42.4	89.4
LC25000-Lung	46.7	57.5	63.2	<u>79.8</u>	92.1
LC25000-Colon	81.3	87.5	92.1	100.0	100.0
CRC100K	59.8	39.9	16.2	57.1	<u>57.4</u>
BACH	29.7	36.3	15.3	<u>42.1</u>	88.8
WSSSLUAD	<u>62.5</u>	60.0	42.1	43.9	85.1
PatchCamylon17	<u>64.4</u>	34.6	58.1	66.2	52.7
Osteo	<u>63.3</u>	54.2	33.6	34.9	78.9
MHIST	<u>50.0</u>	53.8	<u>50.0</u>	<u>50.0</u>	47.4
SICAPv2	<u>41.3</u>	35.6	25.7	26.9	80.9
AGGC2022	<u>51.3</u>	36.8	18.1	27.4	84.4
KIRP	74.6	77.7	59.7	<u>92.2</u>	99.2
PAIP19	54.4	<u>71.8</u>	33.4	43.1	89.4
VALSET_UKK	<u>39.4</u>	30.4	23.1	19.1	87.6
Average	<u>53.7</u>	50.5	38.4	51.8	81.0

Table B.3. Performance comparison of general-purpose and pathology-specific models on OOD patch classification tasks.

	ABMIL	DSMIL	GPT-4o	PRISM	Quilt-LLaVA	PathGen-LLaVA	CPath-Omni (ours)
TCGA-THCA	<u>58.7</u>	59.9	37.6	32.3	29.5	40.5	58.5
TCGA-RCC	95.7	<u>95.3</u>	43.0	50.3	38.6	51.1	94.0
TCGA-ESCA	97.4	97.4	73.7	79.0	63.2	76.3	92.1
TCGA-NSCLC	91.1	87.2	58.8	81.1	58.8	65.5	<u>88.8</u>
TCGA-UCEC	93.4	83.0	42.9	41.4	47.0	47.4	<u>87.8</u>
TCGA-BLCA	60.3	61.7	54.1	<u>68.8</u>	51.4	53.7	70.7
TCGA-BRCA	82.4	<u>86.7</u>	50.6	83.5	48.1	60.5	89.2
TCGA-TGCT	<u>72.6</u>	72.1	42.9	27.4	20.5	39.3	80.9
Average	<u>81.5</u>	80.4	50.5	58.0	44.6	54.3	82.8

Table B.4. Performance comparison of general-purpose and pathology-specific models on WSI classification tasks using balanced accuracy.

Dataset	Classes
PatchCamelyon	<i>'lymph node', 'lymph node metastasis'</i>
NCK-CRC	<i>'Adipose', 'Debris', 'Lymphocytes', 'Mucus', 'Smooth muscle', 'Normal colon mucosa', 'Cancer-associated stroma', 'Colorectal adenocarcinoma epithelium'</i>
LC25000Lung	<i>'Lung adenocarcinoma', 'benign lung tissue', 'lung squamous cell carcinomas'</i>
LC25000Colon	<i>'Colon adenocarcinoma', 'normal colon tissue'</i>
BACH	<i>'Benign tissue', 'In-situ carcinoma', 'Invasive carcinoma', 'Normal tissue'</i>
SICAPv2	<i>'Non-cancerous', 'Atrophic well differentiated and dense glandular regions', 'Cribriform, ill-formed, large-fused and papillary glandular patterns', 'Isolated cells or file of cells, nests of cells without lumina formation and pseudo-rosetting patterns'</i>
Osteo	<i>'Non-tumor', 'Necrotic tumor', 'Viable tumor'</i>
SkinCancer	<i>'Non-tumor chondral tissue', 'Non-tumor dermis', 'Non-tumor elastosis', 'Non-tumor epidermis', 'Non-tumor hair follicle', 'Non-tumor skeletal muscle', 'Non-tumor necrosis', 'Non-tumor nerves', 'Non-tumor sebaceous glands', 'Non-tumor subcutis', 'Non-tumor sweat glands', 'Non-tumor vessel', 'Tumor epithelial basal cell carcinoma', 'Tumor epithelial squamous cell carcinoma', 'Tumor melanoma', 'Tumor naevus'</i>
WSSSLUAD	<i>'tumor', 'normal'</i>

Table B.5. Classes for each dataset in zero-shot image classification. Consistent prompt templates are used for all datasets, including: 'An H&E image of {}', 'This is an image of {} presented in the image', and 'An H&E patch of {}'.

Hyper-parameter	Value
Num GPUs	8
Num epochs	5
Learning rate	3e-5
Per device train batch size	64
Gradient accumulation steps	1
Weight decay	0.1
Warmup steps	300

Table B.6. Hyperparameters used in CPath-CLIP training.

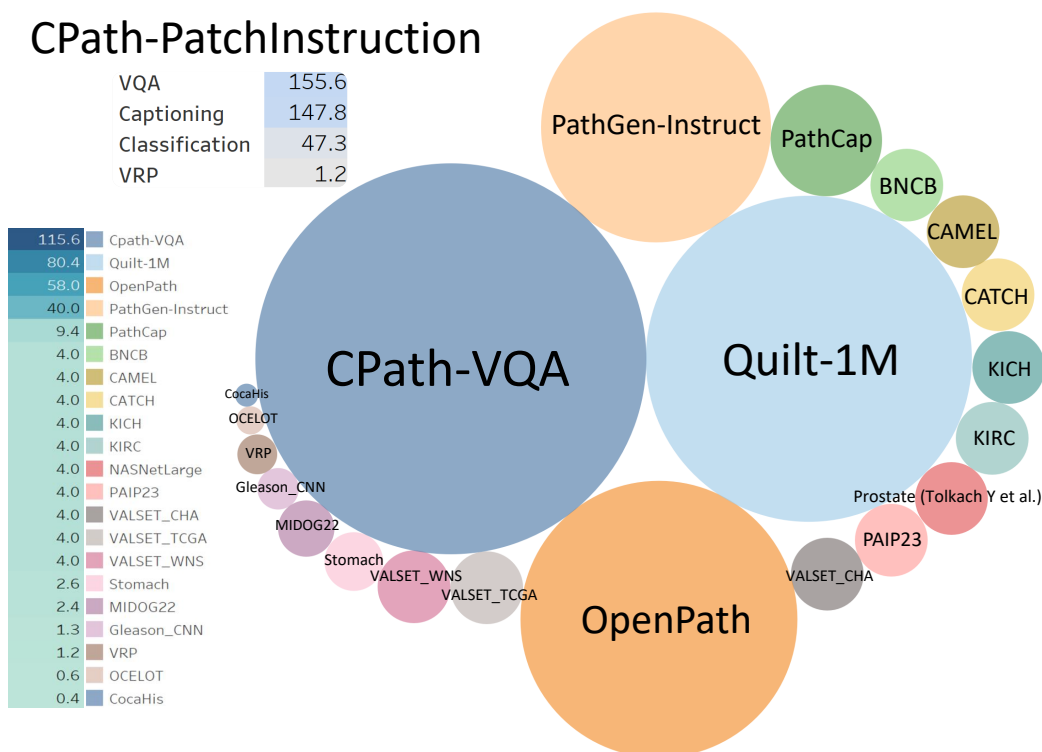


Figure B.3. Visualization of the datasets used in CPath-PatchInstruct, including their quantities (in thousands) and proportional distributions, where larger circles represent higher proportions.

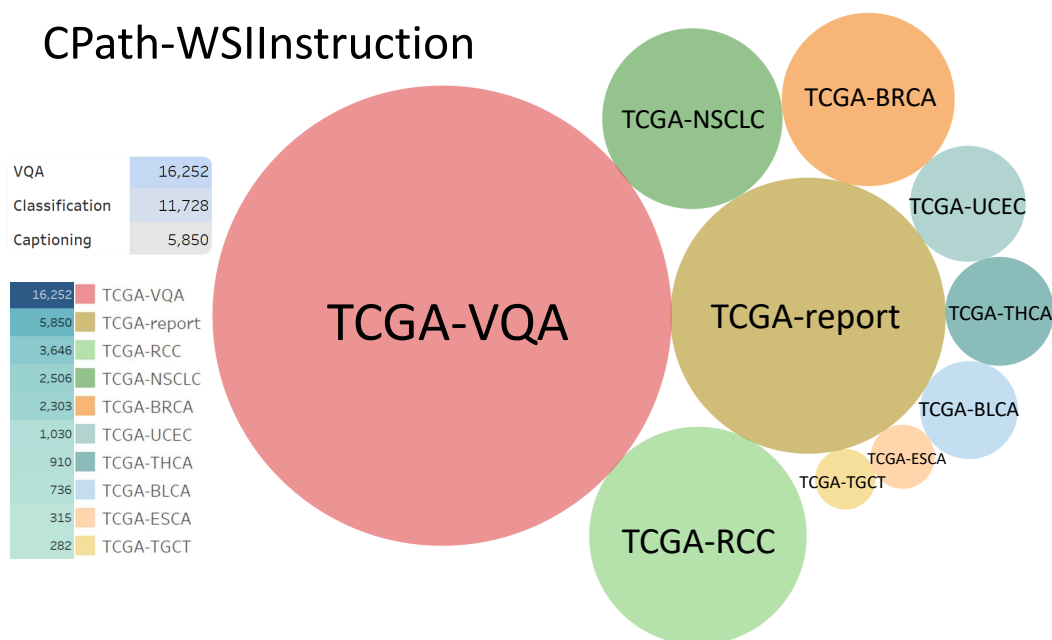


Figure B.4. Visualization of the datasets used in CPath-WSIInstruct, including their quantities and proportional distributions, where larger circles represent higher proportions.

morphology

The image shows clusters of cells with a glandular or acinar arrangement, which is typical for pancreatic tissue. ¹

The cells within these clusters have moderately abundant cytoplasm and round to oval nuclei. Some nuclei appear slightly enlarged and hyperchromatic. ²

There is a noticeable desmoplastic reaction in the surrounding stroma, characterized by dense fibrous tissue. ³

The stroma also contains scattered inflammatory cells, including lymphocytes and possibly plasma cells. ⁴

The nuclei within the cell clusters show mild pleomorphism, with some variation in size and shape. ⁵

There is no significant atypia or prominent nucleoli observed. ⁶

Mitotic figures are not prominently visible in this section. ⁷

There are duct-like structures present, some of which appear dilated. ⁸

The lining epithelium of these ducts does not show significant atypia but may exhibit mild reactive changes. ⁹

No clear evidence of necrosis or hemorrhage is noted in this section. ⁰

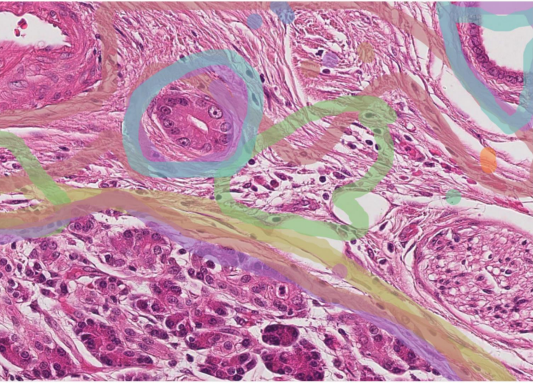
The overall architecture suggests a disruption of normal pancreatic acinar structure. ^q

diagnosis

The presence of desmoplastic stroma, mild nuclear pleomorphism, and disrupted acinar architecture are suggestive of pancreatic adenocarcinoma. ^w

The desmoplastic reaction and architectural distortion are key features supporting the diagnosis of pancreatic adenocarcinoma. ^e

A definitive diagnosis would require correlation with clinical findings, additional histological sections, and possibly immunohistochemical staining. ^t



Info

History

Selection Details

The stroma also contains scattered inflammatory cells, including lymphocytes and possibly plasma cells.

ID 4TePbn8u-a

+

↶

↷

Regions

Relations

Manual

By Time 1:1

The stroma also contains scattered inflammatory cells, including lymphocytes and possibly plasma cells.

Revision

The nuclei within the cell clusters show mild pleomorphism, with some variation in size and shape.

The nuclei in the cell clusters are not found to be abnormal

There is no significant atypia or prominent nucleoli observed.

Revision

Mitotic figures are not prominently visible in this section.

↶

Update

morphology

The histology image of the endometrium shows a dense fibrous stroma with scattered glandular structures. ¹

The glands are irregularly shaped and lined by columnar epithelial cells. ²

Some of the epithelial cells exhibit nuclear atypia. ³

There is a notable presence of stromal fibrosis, characterized by thick bundles of collagen fibers interspersed with spindle-shaped stromal cells. ⁴

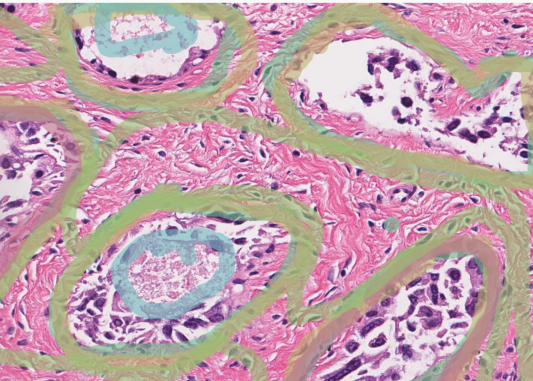
The glandular lumina contain eosinophilic secretions. ⁵

The overall architecture suggests a disorganized pattern. ⁶

diagnosis

The features may be indicative of a pathological process such as endometrial hyperplasia or a neoplastic condition. ⁷

Further clinical correlation and additional diagnostic tests would be necessary for a definitive diagnosis. ⁸



Info

History

Selection Details

revision

Some of the epithelial cells exhibit nuclear atypia.

Revision

There is a notable presence of stromal fibrosis, characterized by thick bundles of collagen fibers interspersed with spindle-shaped stromal cells.

Revision

The glandular lumina contain eosinophilic secretions.

Revision

The overall architecture suggests a disorganized pattern.

A

Update

Figure B.5. Examples of the pathologist annotation interface for visual referring prompting. Pathologists are required to verify whether the given morphology and diagnosis are correct, record "T" or "F" in the bottom-right corner, or modify the original findings as needed. Once all findings are confirmed accurate, they use corresponding colored markers to highlight the regions in the image associated with each finding.

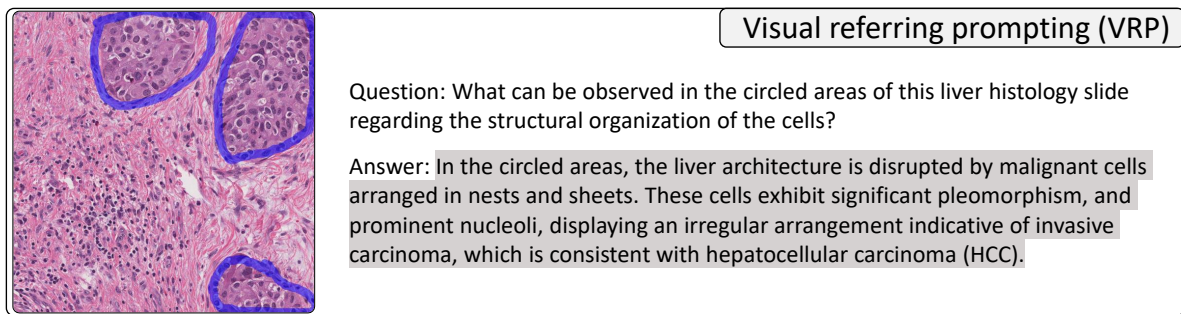


Figure B.6. An example of a constructed visual referring prompting task, where questions are answered based on the highlighted regions.

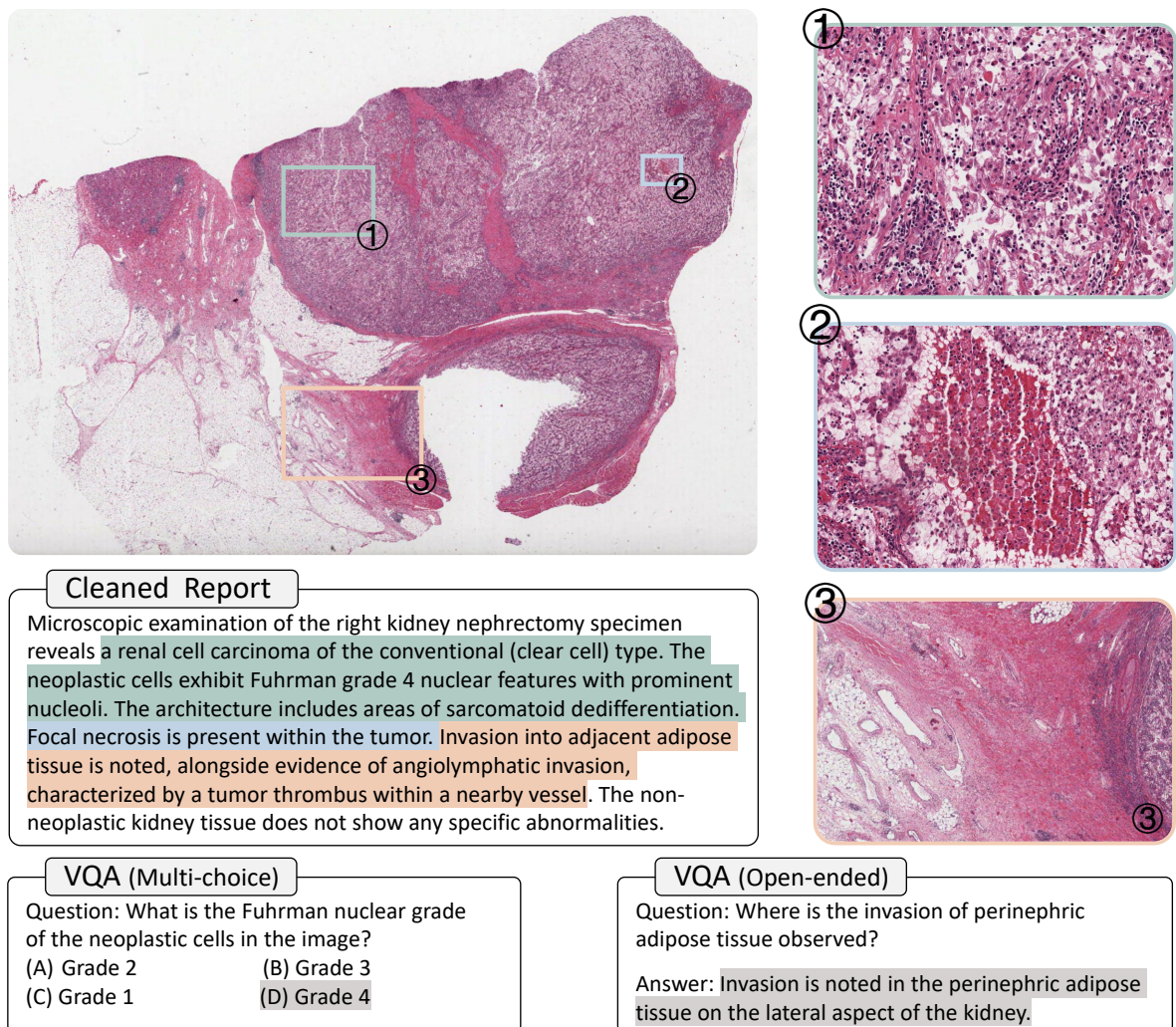


Figure B.7. Examples of WSI Captioning and VQA tasks, where corresponding findings in the captions are highlighted with matching colored boxes in the WSI.

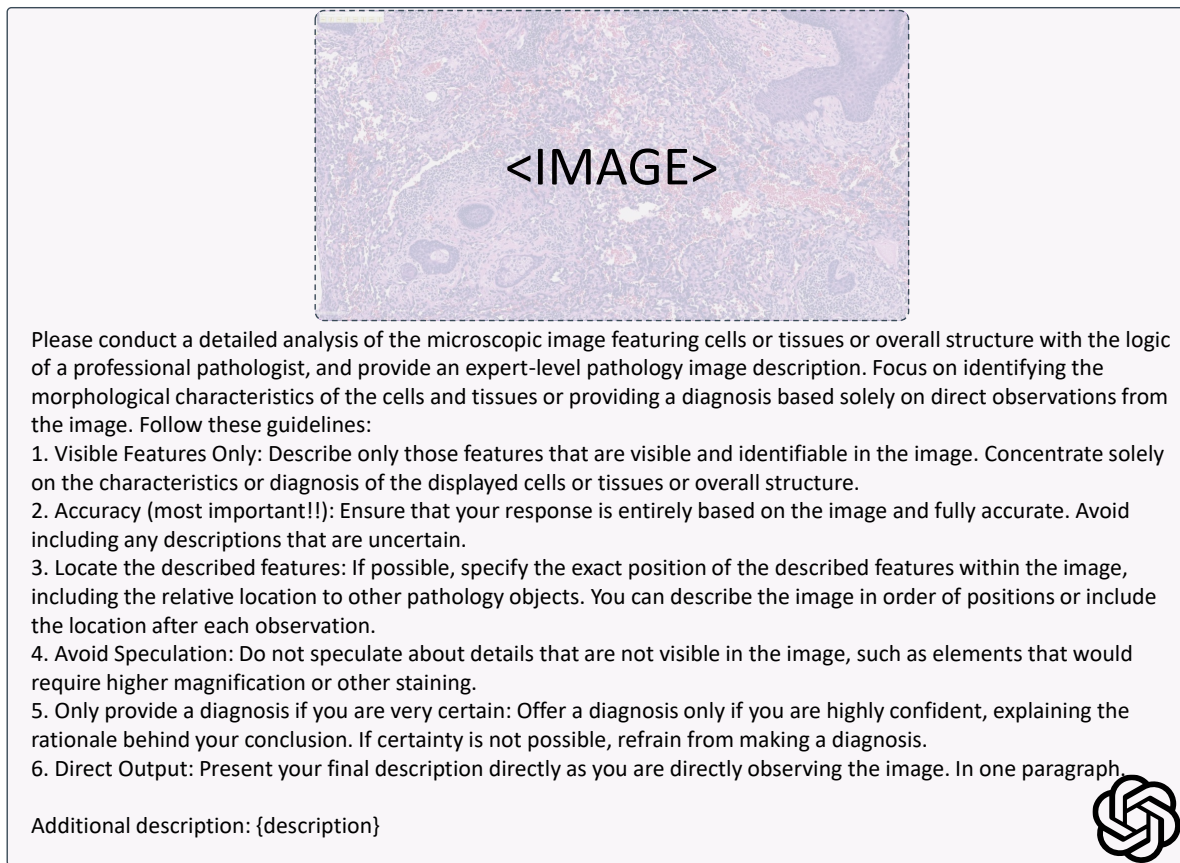


Figure C.8. Prompt for GPT-4o to generate a detailed description for an image based on its original caption.

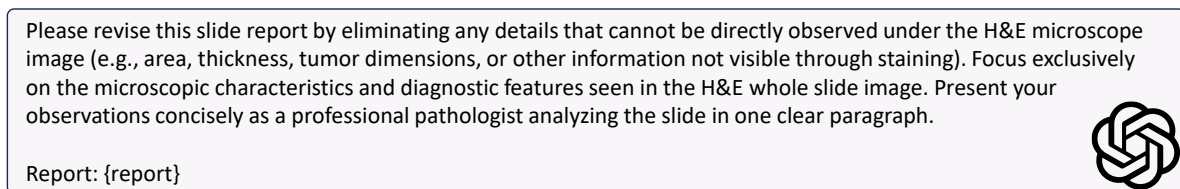


Figure C.9. Prompt for GPT-4o to clean the raw data from the WSI report, transforming it into accurate ground truth WSI report.

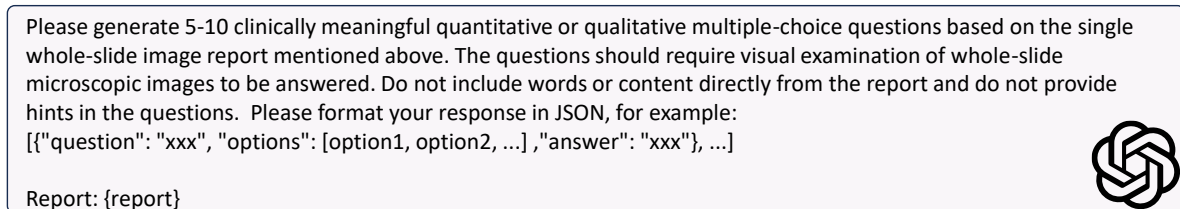


Figure C.10. Prompt for GPT-4o to generate closed-ended VQA based on a given WSI report.

Please generate 5-10 clinically meaningful quantitative or qualitative multi-choice question-and-answer pairs based on the single whole-slide image report mentioned above. The questions should require visual examination of whole-slide microscopic images to be answered. Avoid questions that can be answered solely with medical knowledge without needing to examine the microscopic pathology images. If the content from the report is insufficient, generate fewer questions accordingly. Do not include words or content directly from the report or provide too many hints in the questions. Please format your response in JSON, for example: [{"question": "xxx", "answer": "xxx"}, ...]

Report: {report}



Figure C.11. Prompt for GPT-4o to generate open-ended VQA based on a given WSI report.

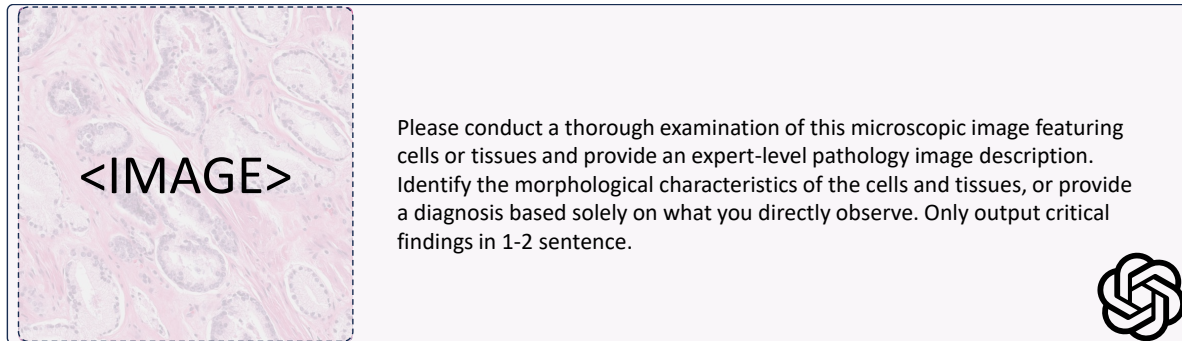


Figure C.12. Prompt for GPT-4o to generate the caption for WSI patch image.

Summarize the following patch captions to generate a concise report for the entire whole slide image, presented in one coherent paragraph.

Patch captions: {captions}



Figure C.13. Prompt for GPT-4o to merge generated patch captions into a comprehensive WSI report.

According to the given pathology whole slide image report, answer the question. Only output the option index.

Report: {report}



Figure C.14. Prompt for GPT-4o to answer the closed-ended question based on the generated WSI report.

According to the given pathology whole slide image report, answer the question in short.

Report: {report}



Figure C.15. Prompt for GPT-4o to answer the open-ended question based on the generated WSI report.

Please determine whether the prediction is correct based on the question and the reference answer. Provide your response, outputting only 'yes' or 'no.'

Question: {question}

Predicted: {predicted answer}

Reference: {reference answer}



Figure C.16. Prompt for GPT-4o to determine whether the predicted answer to the open-ended question is correct.

Hyper-parameter	Value
LLM Model	Qwen2.5-14B-Instruct
Vision Model	CPath-CLIP
Tunable parts	MLP
Vision select layer	-2
Model max length	8192
Image aspect ratio	Square
Image grid pinpoints	None
Patch merge type	Flat
Prompt version	Plain
Num GPUs	8
Num epochs	1
Learning rate	1e-3
Per device train batch size	16
Gradient accum steps	1
Weight decay	0.
Warmup ratio	0.03
Lr scheduler type	cosine

Table B.7. Hyperparameters used in stage 2 training of CPath-Omni (Patch pretraining).

Hyper-parameter	Value
LLM Model	Qwen2.5-14B-Instruct
Vision Model	CPath-CLIP
Tunable parts	CPath-CLIP & MLP & LLM
Vision select layer	-2
Model max length	32768
Image aspect ratio	AnyRes (up to 9 splits)
Image grid pinpoints	(1x1),..., (3x3)
Patch merge type	Spatial unpad
Prompt version	Qwen-1.5
Num GPUs	8
Num epochs	1
Learning rate	1e-5
Per device train batch size	1
Gradient accum steps	8
Weight decay	0.
Warmup ratio	0.03
Lr scheduler type	cosine
Vision tower lr	2e-6

Table B.8. Hyperparameters used in stage 2 training of CPath-Omni (Patch fine-tuning).

Hyper-parameter	Value
LLM Model	Qwen2.5-14B-Instruct
Vision Model	CPath-CLIP
Tunable parts	WSI projector
Vision select layer	-2
Model max length	8192
WSI hidden size	3328
Image aspect ratio	Square
Image grid pinpoints	None
Patch merge type	Flat
Prompt version	Plain
Num GPUs	8
Num epochs	1
Learning rate	5e-6
Per device train batch size	16
Gradient accum steps	1
Weight decay	0.
Warmup ratio	0.1
Lr scheduler type	cosine

Table B.9. Hyperparameters used in stage 3 training of CPath-Omni (WSI pretraining).

Hyper-parameter	Value
LLM Model	Qwen2.5-14B-Instruct
Vision Model	CPath-CLIP
Tunable parts	CPath-CLIP & MLP & WSI projector & LLM
Vision select layer	-2
Model max length	32768
Wsi hidden size	3328
Image aspect ratio	AnyRes (up to 9 splits)
Image grid pinpoints	(1x1),..., (3x3)
Patch merge type	Spatial unpad
Prompt version	Qwen-1.5
Num GPUs	8
Num epochs	5
learning rate	1e-5
Per device train batch size	1
Gradient accum steps	8
Weight decay	0.
Warmup ratio	0.1
Lr scheduler type	cosine
WSI projector lr	1e-5
Vision tower lr	2e-6

Table B.10. Hyperparameters used in stage 4 training of CPath-Omni (mixed patch and WSI fine-tuning).

Dataset	Source Link
TCGA	https://portal.gdc.cancer.gov/
CocaHis	https://portal.gdc.cancer.gov/
BCNB	https://bcnb.grand-challenge.org/
CAMELYON17	https://camelyon17.grand-challenge.org/Data/
MIDOG2022	https://midog.deepmicroscopy.org/download-dataset/
AGGC2022	https://aggc22.grand-challenge.org/
ARCH	https://warwick.ac.uk/fac/cross_fac/tia/data/arch
BACH	https://zenodo.org/records/3632035
CAMEL	https://drive.google.com/open?id=1brr8CnU6ddzAYT157wkDXjbSzoIDF9y
LC2500	https://academictorrents.com/details/7a638ed187a6180fd6e464b3666a6ea0499af4af
MIDOG2021	https://imig.science/midog2021/download-dataset/
OCELOT	https://zenodo.org/record/7844149
Osteo	https://www.cancerimagingarchive.net/collection/osteosarcoma-tumor-assessment/
PAIP2019	https://paip2019.grand-challenge.org/
PAIP2020	https://paip2020.grand-challenge.org/
PAIP2021	https://paip2021.grand-challenge.org/
SICAPv2	https://data.mendeley.com/datasets/9xxm58dvs3/1
CRC-100K	https://zenodo.org/records/1214456
PCam	https://github.com/basveeling/pcam
HistGen	https://github.com/dddavid4real/HistGen
PathGen-Instruct	https://github.com/PathGen-1-6M/PathGen-1.6M
PathCap	https://huggingface.co/datasets/jamessyx/PathCap
OpenPath	https://drive.google.com/drive/folders/1b5UT8BzUphkHZavRG-fmiyY9JWYIWZER
Quilt-1M	https://github.com/wisdomikezogwo/quilt1m
PathMMU	https://huggingface.co/datasets/jamessyx/PathMMU
CocaHis	https://www.sciencedirect.com/science/article/abs/pii/S1746809420305085?via%3Dihub
OCELOT	https://ocelot2023.grand-challenge.org/datasets
Gleason_CNN	https://github.com/eiriniar/gleason_CNN
MIDOG22	https://midog2022.grand-challenge.org
VALSET	https://zenodo.org/records/7548828
WSISEPAIP23	http://www.wisepaip.org/
Prostate (Tolkach Y et al.)	https://zenodo.org/records/3825933
RCdpia (KIRC, KICH, KIRP)	http://39.171.241.18:8888/RCdpia/annotation.php
CATCH	https://www.cancerimagingarchive.net/collection/catch/
SkinCancer	https://heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/data/7QCR8S
MHIST	https://bmirds.github.io/MHIST
WSSSLUAD	https://wsss4luad.grand-challenge.org/
LC25000	https://github.com/tampapath/lung_colon_image_set?tab=readme-ov-file

Table B.11. Datasets used in this study with corresponding access links