

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

<sup>1,2</sup>Yuxuan Sun\*, <sup>2</sup>Yixuan Si\*, <sup>2</sup>Chenglu Zhu, <sup>3</sup>Xuan Gong, <sup>4</sup>Kai Zhang, <sup>1,2</sup>Pingyi Chen,  
<sup>5</sup>Ye Zhang, <sup>1,2</sup>Zhongyi Shui, <sup>2</sup>Tao Lin<sup>†</sup>, <sup>2,6</sup>Lin Yang<sup>†</sup>  
<sup>1</sup>Zhejiang University, USA, <sup>2</sup>Westlake University, <sup>3</sup>Harvard University,  
<sup>4</sup>The Ohio State University, <sup>5</sup>University of Chinese Academy of Sciences,  
<sup>6</sup>Center for Interdisciplinary Research and Innovation, Muyuan

## Abstract

The emergence of large multimodal models (LMMs) has brought significant advancements to pathology. Previous research has primarily focused on separately training patch-level and whole-slide image (WSI)-level models, limiting the integration of learned knowledge across patches and WSIs and resulting in redundant models. In this work, we introduce CPath-Omni, the first 15B parameter LMM that unifies patch and WSI analysis, consolidating a variety of tasks at both levels, including classification, visual question answering, captioning, and visual referring prompting. Extensive experiments demonstrate that CPath-Omni achieves state-of-the-art (SOTA) performance across seven diverse tasks on 39 out of 42 datasets, outperforming or matching task-specific models trained for individual tasks. Additionally, we develop a specialized pathology CLIP-based visual processor for CPath-Omni, CPath-CLIP, which, for the first time, integrates different vision models and incorporates a large language model as a text encoder to build a more powerful CLIP model, which achieves SOTA performance on nine zero-shot and four few-shot datasets. Our findings highlight CPath-Omni’s ability to unify diverse pathology tasks, demonstrating its potential to streamline and advance the field of foundation model in pathology. The code and model are available at [CPath-Omni](#).

## 1. Introduction

Pathology plays a pivotal role in modern medicine, serving as the foundation for diagnosing and understanding diseases [30]. However, pathology requires significant human effort to conduct precise and accurate interpretations of images that can be as large as  $100,000 \times 100,000$  pixels.

\*Equal contribution.

<sup>†</sup>Corresponding author.

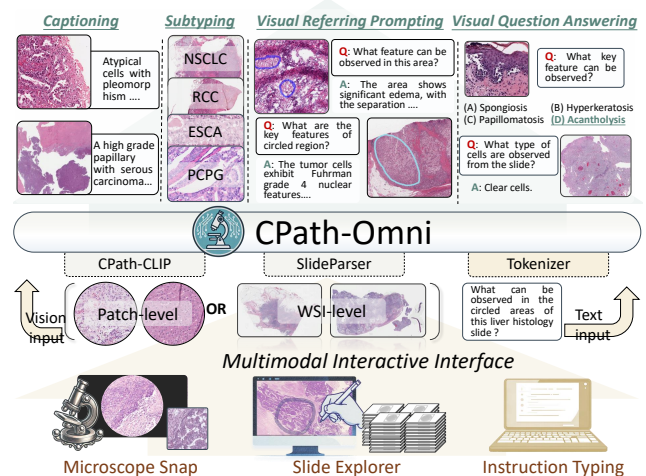


Figure 1. Overview of CPath-Omni’s ability to handle both patch-level and WSI analysis in clinical environments, such as microscope views and scanned WSIs, while supporting various tasks.

In recent years, with advancements in computational power and the digitization of pathology, a wide range of models have been developed to assist pathologists in their diagnostic tasks. At the patch level, these include CLIP [39]-based models like CONCH [34] and PathGen-CLIP [47], DINOv2 [37]-based models like Virchow2 [65] and UNI [12], and LMMs like PathAsst [48], PathGen-LLaVA [47], Quilt-LLaVA [42], and PathChat [35], which support tasks such as multi-turn conversations. At the WSI level, models like Prov-GigaPath [58] and HIPT [11] are developed for WSI classification, while models such as HistGen [15] and WsiCaption [9] are used to generate WSI reports.

In this work, we propose CPath-Omni, a multimodal foundation model designed to unify patch-level and WSI-level analysis. CPath-Omni can perform diverse tasks such as VQA, classification, captioning, and visual referring prompting. By integrating these two levels of analysis and

enabling generalizable task performance, CPath-Omni represents a significant step toward developing a versatile and comprehensive assistive tool for pathologists.

We begin by training a pathology-specific foundation model, CPath-CLIP, to serve as the vision encoder for CPath-Omni. CPath-CLIP is the first model to integrate a large language model (LLM) as the text encoder for CLIP, while incorporating the DINOv2-based pathology vision model Virchow2 alongside the original CLIP-L model as the visual encoder. To train CPath-CLIP, we curate CPath-PatchCaption, a dataset that contains 700,145 high-quality image-caption pairs from diverse sources. Then, we integrate CPath-CLIP into the LLM Qwen2.5-14B [19] to equip it with visual capabilities, creating the CPath-Omni.

The training of CPath-Omni follows four stages to build a unified model capable of handling both patch-level and WSI-level tasks. In the first stage, we pre-align CPath-CLIP with the Qwen2.5-14B LLM using the CPath-PatchCaption dataset. Next, we construct 351,871 instruction tuning samples from four diverse patch-level tasks across 21 datasets, including patch-level classification, VQA, captioning, and visual referring prompting to further finetune CPath-Omni’s patch capabilities. In stage 3, we introduce WSI-related data, including 5,850 cleaned WSI reports, to continue pretraining CPath-Omni, further enhancing its WSI understanding based on the previous stage. In the final stage, we construct 33,830 WSI instruction tuning samples from three tasks across nine datasets, including classification, VQA, and captioning, along with 15% patch-level instruction tuning samples for joint WSI-Patch training. This joint training enables CPath-Omni to seamlessly process both patch and WSI data and enables a wide range of downstream tasks.

Extensive experiments across seven diverse tasks and 42 datasets are conducted to validate the effectiveness of CPath-Omni. With its broad capabilities, CPath-Omni achieves state-of-the-art (SOTA) performance on 39 out of 42 datasets and demonstrates comparable or superior performance to task-specific models. The main contributions of our study are summarized as follows:

- We develop **CPath-CLIP**, the most powerful pathology CLIP model to date, which achieves SOTA results on 9 zero-shot and 4 linear probing classification datasets.
- We introduce **CPath-Omni**, the first unified model capable of handling both patch-level and WSI analysis across diverse tasks, offering promising performance and versatility, and representing an early realization of the “one-for-all” paradigm in computational pathology.
- We curate a diverse and comprehensive training and testing dataset, spanning 7 tasks across 42 datasets, making it the largest and most diverse dataset for training LMMs in pathology. Extensive experiments are conducted on these datasets to confirm CPath-Omni’s significant advancement in pathology foundation models.

## 2. Related Work

**Vision Foundation Models in Pathology.** Recent advances in digital pathology have spurred the development of pathology-specific visual foundation models (VFM), which fall into two main categories. The first is the vision language models like CLIP [39], which employs contrastive learning to align images with textual descriptions, enabling the vision encoder to generate semantically meaningful representations. Researchers have compiled large datasets of image-caption pairs from sources such as PubMed, YouTube, Twitter, and books to train these models. Notable examples include Quilt-Net [21], PLIP [18], PathCLIP [48], PathGen-CLIP [47], and CONCH [34]. The second category focuses on vision-only models, trained through self-supervised learning using vast amounts of patch data extracted from WSIs. These models are typically trained using DINO [8, 37]-like approaches to learn robust visual representations. Prominent models include Lunit [26], UNI [12], Prov-GigaPath [58], and Virchow series [52, 65].

These VFMs have significantly improved image representations, enhancing performance on downstream tasks like patch and WSI classification. CLIP-based models, which are pre-aligned with semantic representations, tend to capture more coarse-grained features and are easier to integrate with LLMs, while DINO-based models excel at fine-grained visual features [25]. In this work, we combine the strengths of both approaches by leveraging OpenAI-CLIP-L and the vision-only Virchow2 as our visual encoder, aligned with the Qwen2-1.5B [59] LLM to enhance visual capabilities and improve alignment with LLM world knowledge.

**Multimodal Generative Foundation Models in Pathology.** The integration of LLMs with vision capabilities has led to advanced LMMs such as GPT-4V [36] and Gemini Pro Vision [49]. These LMMs offer generalized capabilities, which are particularly valuable for pathology, where experts must understand diverse diseases (e.g., lung and liver cancers), work with various tissue types (e.g., prostate, colon), and perform multiple analytical tasks (e.g., tumor classification, survival prediction). Consequently, several pathology-specific LMMs have emerged, including PathAsst [48], PathGen-LLaVA [47], Quilt-LLaVA [41], and PathChat [35]. While these models demonstrate impressive image understanding and conversational abilities, their input size limitations restrict their application primarily to patch-level analyses.

For WSI-specific generative models, several studies have trained smaller multimodal language models. For example, WsiCaption [9] and HistGen [15] focus on WSI caption generation, while WSI-VQA [10] targets WSI-based visual question answering. These earlier approaches primarily utilized publicly available datasets with fewer than 10,000 samples. More recently, PRISM [43], trained on 587,196 internal WSIs, developed a CoCa [60]-like model

capable of zero-shot WSI classification and report generation, representing a significant step toward more generalizable generative foundation models.

However, these models are still limited to processing either patches or WSIs. Inspired by unified medical models like MedDr [17], RadFM [55], and BiomedGPT [62] that unify different medical domains and modalities (although they are also patch-level models), our work represents the first effort in the pathology domain to unify both patch-level and WSI-level analysis within a single LMM.

**Multimodal Datasets in Pathology.** To construct powerful CLIP-based models, substantial high-quality image-caption pairs are essential. At the patch-level, the ARCH [14] contains 8,617 figure-caption pairs related to histology images from medical articles and textbooks. The PathCap [48] offers 207,000 pathology image-caption pairs curated from over 15 million image-text pairs from PubMed and various textbooks. The OpenPath [18] includes 208,414 pairs collected from Twitter posts, while the QUILT-1M [21] contains 768,826 histopathology image-text pairs derived from YouTube video frames. Additionally, to train LMMs, Quilt-Instruct [41] generated 107,131 instruction-tuning samples from YouTube lectures, while PathGen-Instruct [47] created 200K instruction-tuning samples based on synthetic captions from the PathGen-1.6M [47]. In the WSI domain, available data is considerably more limited. WsiCaption [9] and HistGen [15] generate 10,000 and 7,753 WSI captioning samples based on TCGA reports, respectively. More recently, the WSI-VQA [10] expanded these datasets to support VQA tasks for WSIs.

In this paper, we systematically compile these existing datasets and augment them through additional processing. We also incorporate our downstream datasets as the foundation for the training and testing of CPath-Omni.

### 3. Data Preparation

In this section, we introduce the patch-level and WSI-level data required for constructing CPath-Omni.

#### 3.1. Patch Level Dataset

**CPath-PatchCaption:** This dataset is a curated image-caption pairs dataset consisting of 700,145 pairs gathered from various open-source datasets. Specifically, it includes 218,630 pairs from PathCap, 388,932 pairs from Quilt-1M, and 92,583 pairs from OpenPath. We ensured that this caption dataset does not overlap with the test data used in Path-Omni. This dataset serves as a key component for pretraining CPath-CLIP and the stage 1 pretraining of CPath-Omni.

**CPath-PatchInstruction:** CPath-PatchInstruction is a diverse dataset comprising 351,871 samples across captioning, VQA, classification, and visual referring prompting tasks. Of these, 147,843 highest-quality samples are manually curated from CPath-PatchCaption, with their captions

and images further enhanced through GPT-4 to produce more comprehensive and detailed descriptions. Additionally, 40,000 examples are drawn from PathInstruct-200K, a multimodal, multi-turn conversational pathology dataset. For the classification data, samples are collected from various public classification datasets, including VALSET-TCGA, VALSET-WNS, VALSET-CHA [50], Stomach, KIRC, CocaHis [46], PAIP23, BNCB [56], CATCH [54], PAIP21, MIDOG22 [4], KICH, CAMEL [57], Gleason-CNN [3], OCELOT [40], and Prostate (Tolkach Y et al.) [66]. From each dataset, we randomly sample up to 5,000 samples, converting 80% into VQA format for CPath-Instruction training while reserving the remaining 20% for validation and testing. To further enhance the model’s capabilities and interpretability, we invite expert pathologists to annotate 1,300 high-resolution images with captions selected from the TCGA dataset. These captions are first processed by GPT-4 to generate preliminary findings, which are then meticulously reviewed, supplemented, and refined by pathologists to ensure accuracy. Most importantly, corresponding regions for each pathology finding are highlighted in the images, creating visually grounded prompting data. Of these annotated samples, 1,200 are designated for training, while 100 are reserved for validation and testing. The entire CPath-PathInstruction dataset is utilized for stage 2 training of CPath-Omni. For further details on dataset construction and annotation, please refer to the Appendix.

#### 3.2. WSI Level Dataset

**CPath-WSIInstruction:** CPath-WSIInstruction encompasses captioning, VQA, and classification data at the WSI level. The dataset includes 7,312 WSI-level captioning examples sourced from HistGen. To ensure that the validation and test sets do not overlap with the WSI classification data, we re-divide the data into training, validation, and test sets with an 8:1:1 ratio. The training set is incorporated into CPath-WSIInstruction. For the VQA component, we further generate a WSI VQA dataset by prompting GPT-4 based on the divided WSI captions. Specific prompts are detailed in the Appendix. For classification, we compile subtype data for 8 TCGA subtyping tasks, including RCC, NSCLC, BRCA, UCEC, THCA, ESCA, BLCA, and TGCT. 80% of this data is transformed into the QA format that CPath-Omni can accept for training, while the remaining 20% is split equally for validation and testing.

### 4. The Proposed CPath-Omni

As shown in Fig. 1, CPath-Omni’s architecture includes two vision components, CPath-CLIP for patch-level and SlideParser for WSI-level processing, along with an LLM (Qwen2.5-14B). Patch and WSI inputs are processed through separate branches, generating patch-level or WSI-level visual tokens that are fed into the LLM. In this section,

we detail the construction of CPath-CLIP and SlideParser.

#### 4.1. CPath-CLIP

Pretrained CLIP models are commonly used with LLMs in general-domain LMMs, as they are well-aligned with the text space and are easier to integrate with LMMs. However, pathology images differ significantly from natural images, creating a domain gap that limits CLIP performance in pathology tasks. Therefore, a pathology-specific CLIP-based model should be constructed to improve image understanding and enhance LMM performance in this domain.

One challenge with CLIP models is their focus on coarse-grained semantic information, which can cause the loss of fine-grained details critical for pathology diagnoses, such as chromatin structure or mitosis. To address this, we integrate the pretrained Virchow2 model, trained on 3 million WSIs using DINOv2-based vision pretraining, providing strong visual representations. Simultaneously, we retain OpenAI’s CLIP vision component to preserve strong semantic features. As illustrated in the left part of Fig. 2, we feed the image into both models and concatenate their feature outputs. In addition, to enhance alignment between the vision model and the LLM, we replace the GPT-2 model from the original CLIP architecture with the text embedding version of Qwen2-1.5B [6]. With its larger size and more comprehensive training corpus, Qwen2-1.5B introduces stronger world knowledge, substantially improving semantic alignment between vision and language components. Specifically, we train CPath-CLIP using the CPath-PatchCaption dataset based on OpenCLIP framework [23].

When used for patch processing in CPath-Omni, CPath-CLIP adopts the AnyRes strategy from LLaVA-Next [33] to handle higher resolutions. Each image is split into a  $3 \times 3$  grid of sub-patches, which are encoded by CPath-CLIP to extract features. These features are then processed through a two-layer MLP before being input into the LLM.

#### 4.2. SlideParser

SlideParser serves as the core component for handling WSI inputs, particularly for gigapixel images up to  $100,000 \times 100,000$  pixels. To manage this, we first split the WSI into multiple  $2048 \times 2048$  image regions. As pathologists often require both global and local context during analysis, we implement a multi-scale region encoding approach. As shown in Fig. 2, each  $2048 \times 2048$  region is further subdivided into three scales: 16 tiles of  $512 \times 512$ , 4 tiles of  $1024 \times 1024$ , and 1 tile of  $2048 \times 2048$ . These tiles are encoded by CPath-CLIP to generate scale-specific features, which are aggregated via average pooling to produce a multi-scale feature representation for the  $2048 \times 2048$  region. This approach captures detail at multiple scales while efficiently reducing the number of image tokens fed into the LLM.

Since WSIs can vary greatly in size—from dozens to

thousands of patches—this variability can cause instability during LMM training. To address this, we introduce a token compression layer that standardizes the input by reducing WSI tokens to a fixed length. Specifically, we adopt the CoCa approach [60], using 1152 query tokens through multi-head attention to query the patch tokens, resulting in a unified output of 1152 tokens for the LLM.

#### 4.3. The Training of CPath-Omni

The CPath-Omni model architecture adopts the LLaVA-NEXT framework but undergoes a four-stage training process: patch-based pretraining, patch-based finetuning, WSI-based pretraining, and mixed patch-WSI training.

**Stage 1:** This initial phase focuses on aligning the feature spaces between CPath-CLIP and the LLM. Only the two-layer MLP connecting CPath-CLIP to the LLM is trained using the CPath-PatchCaption dataset, enabling effective pre-alignment of visual and language features.

**Stage 2:** All model parameters are unfrozen for finetuning using the CPath-PatchInstruct dataset. Building on Stage 1’s alignment, this phase enables CPath-Omni to learn a variety of tasks, including VQA, image classification, captioning, and pathology-related knowledge.

**Stage 3:** Training in this stage utilizes exclusively WSI report data. Only SlideParser is unfrozen to align WSI features with the pathology-related knowledge previously acquired by the LLM during patch-based training.

**Stage 4:** The final stage introduces mixed training with 15% randomly sampled CPath-PathInstruct and CPath-WSIInstruct datasets. By this point, CPath-Omni has gained a strong understanding of pathology from previous stages, enabling effective transfer of patch-based knowledge to WSI tasks, despite limited WSI data availability.

After completing these four stages, CPath-Omni is comprehensively equipped to handle both patch-based and WSI analysis across a variety of downstream tasks.

### 5. Experiments

We conduct extensive experiments to evaluate CPath-Omni’s universal task-solving capabilities. At the patch level, we evaluate across 32 subsets, including tasks such as VQA, classification, captioning, and visual referring prompting (VPR). For the WSI level, we evaluated 10 subsets focusing on WSI VQA, classification, and captioning.

#### 5.1. Benchmarking CPath-CLIP

In our experiments, we evaluate the image-text alignment and feature extraction capabilities of CPath-CLIP through zero-shot classification and few-shot linear probing. For zero-shot classification, we utilize datasets including Patch-Camelyon (Pcam) [51], CRC-100K [27], SICAPv2 [44], BACH [1], Osteo [2], SkinCancer [29], WSSSLUAD [16], LC-Lung, and LC-Colon [7]. We benchmark our model’s

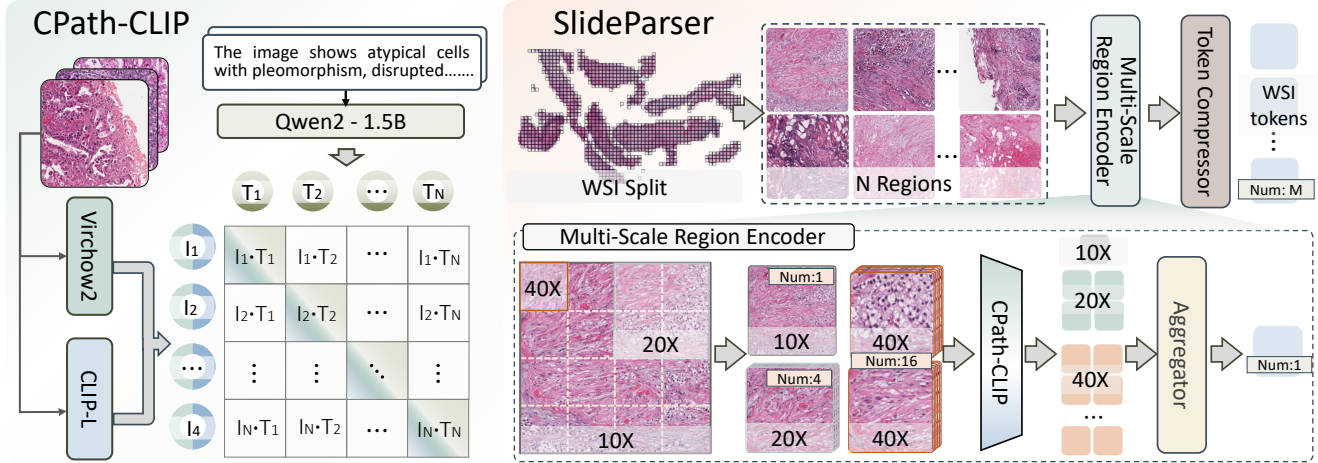


Figure 2. Overview of two key vision components of CPath-Omni: the patch-level model, CPath-CLIP, and the WSI model, SlideParser.

Model	LC-Lung	LC-Colon	CRC100K	SkinCancer	Pcam	BACH	Osteo	WSSSLUAD	SICAPv2	Average
OpenAI-CLIP-L	70.4	81.1	40.3	19.4	55.5	34.3	53.9	81.2	25.4	51.3
PLIP	87.9	90.2	52.8	42.5	51.8	34.3	52.9	73.1	42.5	58.6
QuiltNet	80.0	91.0	49.5	46.4	58.7	43.8	53.8	70.5	37.3	58.9
PathCLIP	88.9	94.3	55.3	35.1	72.5	46.8	69.2	<u>85.1</u>	48.3	66.2
KEP	91.6	98.9	64.4	46.3	68.4	55.0	47.8	79.9	33.9	65.1
BiomedCLIP	48.8	94.3	29.9	31.7	84.0	39.8	36.7	73.7	32.2	52.9
PathGen-CLIP-L	<u>89.8</u>	<u>99.3</u>	<b>78.0</b>	<u>70.6</u>	<u>88.2</u>	<u>71.5</u>	<u>74.6</u>	82.2	<b>63.5</b>	<u>79.7</u>
CPath-CLIP	<b>97.1</b>	<b>100.0</b>	<b>78.0</b>	<b>74.2</b>	<b>95.9</b>	<b>72.3</b>	<b>80.7</b>	<b>87.1</b>	<u>63.1</u>	<b>83.2</b>

Table 1. Zero-shot classification comparison of various CLIP models on different pathology image classification datasets with accuracy (%). The best performance is highlighted in **bold**, while the second-best is underlined.

performance against OpenAI-CLIP-L [38], PLIP [18], QuiltNet [22], PathCLIP [48], BiomedCLIP [63], KEP [64], and the SOTA PathGen-CLIP-L [47].

**Results: CPath-CLIP achieved superior performance across most datasets.** As shown in Tab. 1, CPath-CLIP notably surpasses the current SOTA model, PathGen-CLIP-L by a significant margin on the Osteo, Pcam, and LC-Lung datasets, with improvements of 6.1%, 7.7% and 7.3%, respectively. Additionally, it demonstrated significant advantages over other models, underscoring CPath-CLIP’s enhanced image-text feature alignment capabilities driven by its stronger vision and language model integration.

For few-shot linear probing, we added a fully connected layer to extracted feature representations from four datasets: LC-Colon, Camelyon17, LC-Lung, and WSSSLUAD, using training sizes of 2, 8, 16, 32, 64, and 128 shots. Each size was randomly sampled 10 times and evaluated over 10 runs per configuration. Box plots are utilized to illustrate the variability and robustness of the model’s performance.

**Results: CPath-CLIP consistently outperformed previous models.** As shown in Fig. 3, CPath-CLIP demonstrated rapid improvement with minimal data, reaching 95% accuracy on CRC and LC-Lung using only 2 shots. In contrast,

other models achieved less than 91% accuracy on CRC and approximately 92% on LC-Lung.

*We hypothesize that using advanced LLMs as CLIP’s text encoder provides superior world knowledge compared to previous approaches using BERT or GPT-2. When combined with pathology-specific vision model Virchow2, this integration can achieve faster and more effective alignment.* Unlike general CLIP-based models that require 400 million to billions of training samples, CPath-CLIP was trained with only 700K samples, which holds great potential to redefine future CLIP training paradigms.

## 5.2. Benchmarking CPath-Omni at Patch-Level

At the patch level, we benchmark CPath-Omni against SOTA models, both general-purpose and domain-specific.

We begin by evaluating the performance of various LLMs on VQA using the PathMMU dataset, the largest pathology-specific VQA dataset, which also includes pathologist scores. We compare general-purpose models such as InstructBLIP-FLAN-T5 XXL [13], LLaVA-1.5-13B [32], Qwen-VL-MAX [5], Gemini Pro Vision [49], and GPT-4V [36], alongside domain-specific models such as LLaVA-Med, Quilt-LLaVA, and PathGen-LLaVA.

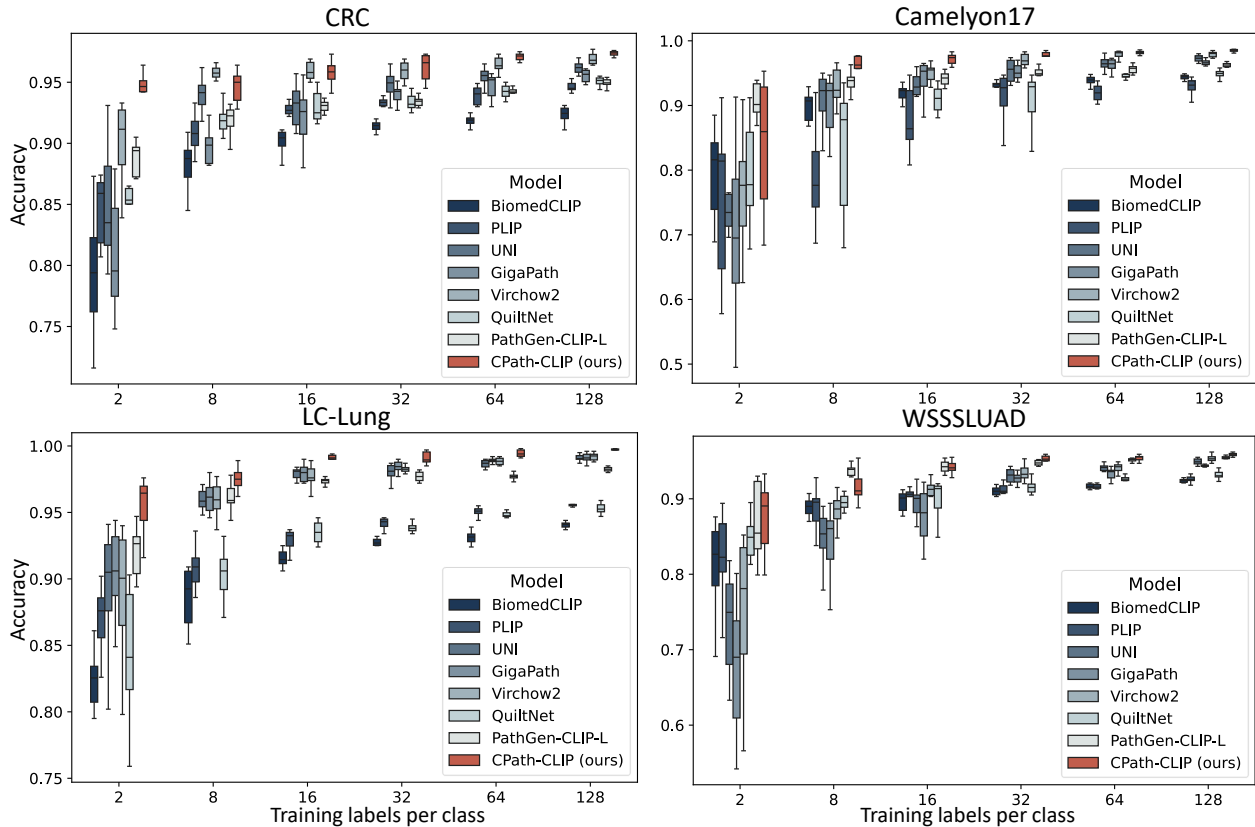


Figure 3. Comparison of few-shot classification accuracy (%) via linear probing across various datasets using different CLIP models.

**Results:** *CPath-Omni significantly outperforms both the latest pathology-specific model, PathGen-LLaVA, and advanced general-purpose models like GPT-4V, even surpassing human-level performance.* As shown in Tab. 2, CPath-Omni exceeds the current SOTA model, PathGen-LLaVA, by 13.8%, with a particularly notable 30.1% improvement in PathCLS [61] performance. Moreover, it slightly surpasses the 71.8% accuracy achieved by human pathologists, by 0.6%. We attribute this to the fact that pathologists annotating the PathMMU dataset may not be experts in all disease categories, whereas foundation models like CPath-Omni can learn generalized knowledge across diverse medical fields. This highlights that CPath-Omni has the promising potential of LMMs to offer valuable assistance to clinicians in real-world settings.

For the classification task, we evaluate performance across 30 classification datasets, including 16 in-distribution (ID) datasets—where the training data is part of CPath-Omni’s training set—such as VALSET-TCGA, VALSET-CHA, VALSET-WNS [50], Stomach, KIRC, CocaHis [46], WISEPAIP23, BNCB [56], CATCH [54], PAIP21, MIDOG22 [4], KICH, CAMEL [57], Gleason-CNN [3], OCELOT [40], and Prostate (Tolkach Y et al.) [66]. Additionally, we use 14 out-of-distribution

(OOD) datasets that were not included in CPath-Omni’s training data, such as AGGC2022 [20], KIRC, PAIP19 [28], VALSE-TUKK, as well as 10 datasets from PathCLS within the PathMMU dataset, namely: Skincancer, LC25000-Lung, LC25000-Colon, CRC-100K [27], BACH, WSSSLUAD, PatchCamlyon17, Osteo, MHIST [53], and SICAPv2 [45]. For all 30 datasets, we compare CPath-Omni with the state-of-the-art models GPT-4o and Gemini-1.5-Pro on their respective test sets. For the ID datasets, we also perform task-specific training with Virchow2, the previous SOTA vision-only model, to facilitate direct comparisons with CPath-Omni.

**Results:** *CPath-Omni significantly outperforms both GPT-4o and Gemini-1.5-Pro, and even achieves performance comparable to a task-specific fine-tuned version of Virchow2 on each individual dataset.* As shown in Fig. 4, the radar chart for CPath-Omni closely mirrors that of Virchow2, with performance varying across datasets. However, CPath-Omni holds a slight advantage in average performance. In contrast, for general-purpose models, CPath-Omni consistently outperforms GPT-4o, even on OOD datasets, demonstrating its superior capability and generalization compared to the strongest general-purpose models. Interestingly, when examining the datasets from the

	Test Overall		PubMed		SocialPath		EduContent		Atlas		PathCLS	
	Tiny (1156)	ALL (9677)	Tiny (281)	ALL (3068)	Tiny (235)	All (1855)	Tiny (255)	All (1938)	Tiny (208)	ALL (1007)	Tiny (177)	ALL (1809)
Expert performance	71.8	-	72.9	-	71.5	-	69.0	-	68.3	-	78.9	-
<b>General Large Multimodal Models</b>												
InstructBLIP-FLAN-T5-XXL	34.3	33.9	39.1	37.2	33.6	34.3	34.5	36.0	38.5	39.3	22.6	22.7
LLaVA-1.5-13B	38.8	37.6	44.5	41.0	40.4	40.4	34.1	39.4	47.1	44.3	24.9	23.5
Qwen-VL-MAX	49.2	45.9	53.0	50.9	53.6	49.3	52.2	47.9	<u>51.4</u>	49.8	30.5	29.6
Gemini Pro Vision	42.8	42.7	43.8	44.9	42.4	42.0	43.5	43.7	49.5	49.4	32.8	<u>34.7</u>
GPT-4V-1106	<u>53.9</u>	<u>49.8</u>	<u>59.4</u>	<u>53.5</u>	<u>58.7</u>	<u>53.9</u>	<u>60.4</u>	<u>53.6</u>	48.1	<u>52.8</u>	<u>36.2</u>	33.8
<b>Pathology-specific Large Multimodal Models</b>												
LLaVA-Med	25.3	26.2	28.5	27.7	28.9	27.3	22.7	27.2	22.6	30.7	22.6	20.3
Quilt-LLaVA	45.6	41.5	47.3	42.6	46.4	46.6	51.8	45.3	46.2	42.7	32.2	29.2
PathGen-LLaVA	60.1	58.4	60.1	60.1	60.9	58.8	60.8	60.7	63.5	64.9	54.2	48.9
CPath-Omni	<b>72.4</b>	<b>72.2</b>	<b>74.0</b>	<b>69.9</b>	<b>76.6</b>	<b>71.8</b>	<b>69.8</b>	<b>70.6</b>	<b>65.9</b>	<b>70.6</b>	<b>75.7</b>	<b>79.0</b>

Table 2. Overall results of models on the PathMMU test set. The best-performing LMM in each subset for general and pathology domain LMMs is **in-bold**, and the top-performing LMM is underlined.

PathCLS branch, CPath-Omni’s overall OOD performance is strikingly close to that of CLIP-based models in a zero-shot setting (refer to Tab. 1). This suggests that CPath-Omni effectively harnesses the power of vision models within its multimodal framework, achieving zero-shot visual classification performance comparable to that of CLIP.

For visual referring prompting and patch captioning, we compare CPath-Omni with domain-specific models including PathGen-LLaVA, Quilt-LLaVA, and LLaVA-Med. We evaluate using 50 manually annotated images and through both GPT-4o assessment (comparing outputs to pathologist-provided ground truth) and human evaluation.

**Results: CPath-Omni significantly outperforms the comparison models in both GPT-4o and human evaluations, with the lowest win rate reaching 84% in the VPR task when compared to Quilt-LLaVA.** Interestingly, CPath-Omni achieves an even higher win rate against PathGen-LLaVA in the VPR task, despite PathGen-LLaVA being a stronger overall model. We hypothesize that this performance difference arises from the differences in training data: Quilt-LLaVA is trained on videos from YouTube instructors, which may include scenarios resembling visual referring prompting, whereas PathGen-LLaVA is primarily trained on synthetic TCGA data, which lacks such data.

### 5.3. Benchmarking CPath-Omni at WSI-Level

In WSI tasks, we benchmark CPath-Omni against both general-purpose models and task-specific fine-tuned models of 3 WSI tasks across 10 datasets.

For the classification task, we evaluate eight TCGA subtyping datasets, including RCC (Kidney Chromophobe, Kidney Renal Clear Cell Carcinoma, Kidney Renal Papillary Cell Carcinoma), NSCLC (Lung Adenocarcinoma, Lung Squamous Cell Carcinoma), BRCA (Invasive Ductal

CPath-Omni VS.	GPT-4o-eval		Human-eval	
	VPR	Captioning	VPR	Captioning
PathGen-LLaVA	96%	82%	98%	80%
Quilt-LLaVA	84%	96%	90%	92%
LLaVA-Med	96%	98%	100%	100%

Table 3. Comparison of CPath-Omni performance on Patch-level VPR and Captioning tasks with GPT-4o-eval and Human-eval across different models. The number represent the percentage of CPath-Omni’s responses considered superior.

Carcinoma, Invasive Lobular Carcinoma), UCEC (Cystic Mucinous and Serous Neoplasms, Adenomas and Adenocarcinomas), THCA (Papillary Adenocarcinoma, Papillary Carcinoma Columnar Cell, Papillary Carcinoma Follicular Variant), ESCA (Adenomas and Adenocarcinomas, Squamous Cell Neoplasms), BLCA (Transitional Cell Carcinoma, Papillary Transitional Cell Carcinoma), and TGCT (Non-seminoma, Mixed-seminoma, Seminoma). For more detailed information, please refer to the Appendix.

Given that these subtyping tasks are incorporated in CPath-Omni’s training, we test on the held-out test set and compare CPath-Omni’s performance against task-specific models, including ABMIL [24], DSMIL [31], a pathology CoCa-style pre-trained model (PRISM), and GPT-4o.

For ABMIL and DSMIL, we use features extracted from WSI patches via CPath-CLIP as input and train the models for 20 epochs, selecting the best checkpoints from the evaluation set for testing on the test set. In PRISM, we use the prompts from the PRISM paper for the eight subtyping tasks it covers. For tasks not included in PRISM, we use the TCGA subtyping classification names as prompts. For GPT-4o, since it does not directly support WSI diagno-

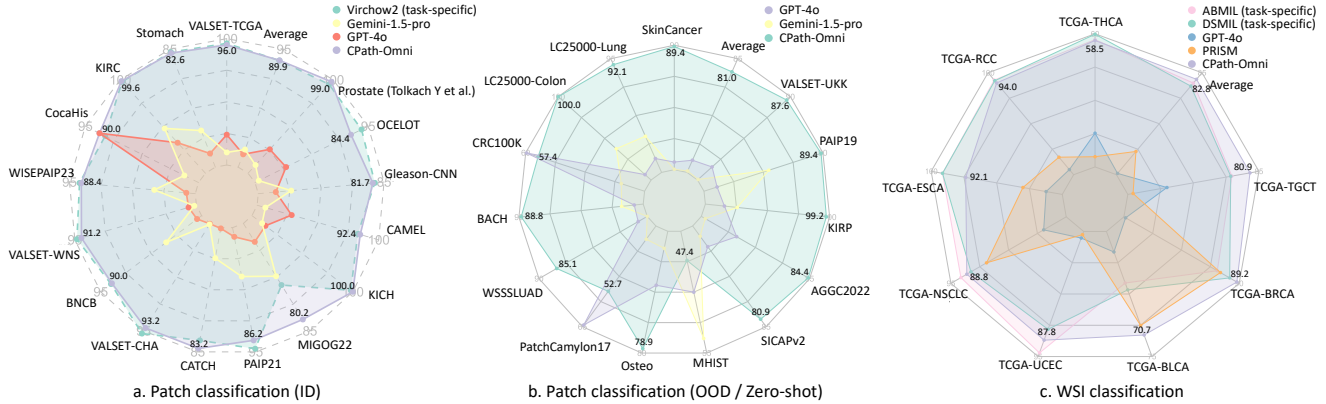


Figure 4. Radar plot visualization of CPath-Omni’s performance on patch and WSI classification tasks: (a) patch-level performance under ID conditions, (b) patch classification performance under OOD/zero-shot conditions, and (c) whole-slide image (WSI) performance.

sis, we first split the WSI into  $4096 \times 4096$  patches at 20X magnification. GPT-4o then generates descriptions for each patch, and these are merged to form a report for the WSI. This report is subsequently used as input to prompt GPT-4o for classification of the WSI.

**Results:** *CPath-Omni significantly outperforms GPT-4o and the pathology-specific foundation model PRISM, while demonstrating comparable or superior performance to task-specific fine-tuned models such as ABMIL and DSMIL.* As shown in Fig. 4, CPath-Omni surpasses ABMIL and DSMIL on three tasks—TCGA-BLCA, TCGA-BRCA, and TCGA-TGCT—and shows a slight overall performance advantage over these models. This suggests that, in the future, a unified framework for WSI classification could achieve the performance of specialized models without the need for task-specific fine-tuning.

For WSI report generation, we compare CPath-Omni with task-specific models, including WsiCaption and HistGen, as well as the general-purpose model PRISM (also supports report generation) and GPT-4o (using the aforementioned method to generate WSI reports). Performance is evaluated using BLEU 1-4 and ROUGE-L scores. For VQA, since the first three models do not support it, we focus the comparison on CPath-Omni and GPT-4o, using the WSI reports generated by GPT-4o as context for answering questions. For closed-ended questions, accuracy is computed, while for open-ended questions, due to the brevity of answers, even small variations or synonyms can cause significant fluctuations in metrics like BLEU and recall. Therefore, we prompt GPT-4o to reference standard answers to evaluate response accuracy instead of BLEU and recall. Note that WSI report generation and VQA tasks do not overlap with data used for WSI classification training.

**Results:** *CPath-Omni achieves SOTA performance in both WSI report generation and WSI VQA tasks, as shown in Tab. 4.* CPath-Omni slightly outperforms the previous

SOTA model, HistGen. Note that PRISM tends to generate very short reports (often only a few words expressing the classification), which results in relatively lower performance metrics. In the WSI VQA task, CPath-Omni significantly outperforms GPT-4o, with performance metrics almost doubling: 67.3% vs. 20.5% for open-ended questions and 70.8% vs. 35.5% for closed-ended questions.

Model	Report Generation					VQA	
	BLEU <sub>1</sub>	BLEU <sub>2</sub>	BLEU <sub>3</sub>	BLEU <sub>4</sub>	ROUGE <sub>L</sub>	Open	Closed
WsiCaption	21.8	13.7	8.6	6.4	25.1	-	-
HistGen	31.8	19.7	12.7	8.4	25.4	-	-
PRISM	0.0	0.0	0.0	0.0	8.0	-	-
GPT-4o	15.8	6.2	2.4	0.1	12.8	20.5	35.5
CPath-Omni	<b>33.7</b>	<b>20.1</b>	<b>12.9</b>	<b>8.7</b>	<b>25.6</b>	<b>67.3</b>	<b>70.8</b>

Table 4. Comparison of CPath-Omni’s performance with task-specific and general models on WSI captioning and VQA tasks.

## 6. Conclusion

In this paper, we present CPath-Omni, a versatile foundational multimodal model designed to tackle both patch-level and WSI-level tasks, spanning captioning, classification, VQA, and visual referring prompting. CPath-Omni’s approach enables unified patch and WSI-level training across diverse datasets, allowing knowledge learned from the patch level to simultaneously enhance WSI performance, even trained on a fraction of the data compared to patch-level datasets. Extensive experiments demonstrate that CPath-Omni achieves superior performance across both patch and WSI-level tasks, comparable to or even outperforming task-specific models and significantly surpassing pretrained general-purpose foundation models like PRISM and GPT-4o. These results highlight the potential of LMMs like CPath-Omni to serve as a “one-for-all” solution, advancing the next generation of pathology-specific LMMs.

## 7. Acknowledgements

This study was partially supported by Zhejiang Provincial Natural Science Foundation of China (Grant No.XHD23F0201), the National Natural Science Foundation of China (Grant No.92270108), foundation of Muyuan Laboratory (Program ID: 14106022401,14106022402), the Research Center for Industries of the Future (RCIF) at Westlake University, and the Westlake Education Foundation.

## References

- [1] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019. 4
- [2] Harish Babu Arunachalam, Rashika Mishra, Ovidiu Daescu, Kevin Cederberg, Dinesh Rakheja, Anita Sengupta, David Leonard, Rami Hallac, and Patrick Leavey. Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *PloS one*, 14(4):e0210706, 2019. 4
- [3] Eirini Arvaniti, Kim Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter Wild, Jan Hendrik Rüschoff, and Manfred Claassen. Replication Data for: Automated Gleason grading of prostate cancer tissue microarrays via deep learning., 2018. 3, 6
- [4] Marc Aubreville, Nikolas Stathonikos, Taryn A Donovan, Robert Klopffleisch, Jonas Ammeling, Jonathan Ganz, Frauke Wilm, Mitko Veta, Samir Jabari, Markus Eckstein, et al. Domain generalization across tumor types, laboratories, and species—insights from the 2022 edition of the mitosis domain generalization challenge. *Medical Image Analysis*, 94:103155, 2024. 3, 6
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 5
- [6] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dmzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024. 4
- [7] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Masstorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019. 4
- [8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2
- [9] Pingyi Chen, Honglin Li, Chenglu Zhu, Sunyi Zheng, Zhongyi Shui, and Lin Yang. Wsicaption: Multiple instance generation of pathology reports for gigapixel whole-slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 546–556. Springer, 2024. 1, 2, 3
- [10] Pingyi Chen, Chenglu Zhu, Sunyi Zheng, Honglin Li, and Lin Yang. Wsi-vqa: Interpreting whole slide images by generative visual question answering. In *European Conference on Computer Vision*, pages 401–417. Springer, 2025. 2, 3
- [11] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1
- [12] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 1, 2
- [13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 5
- [14] Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16549–16559, 2021. 3
- [15] Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Lian-sheng Wang, and Hao Chen. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. *arXiv preprint arXiv:2403.05396*, 2024. 1, 2, 3
- [16] Chu Han, Xipeng Pan, Lixu Yan, Huan Lin, Bingbing Li, Su Yao, Shanshan Lv, Zhenwei Shi, Jinhai Mai, Jiatai Lin, et al. Wss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. *arXiv preprint arXiv:2204.06455*, 2022. 4
- [17] Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv e-prints*, pages arXiv–2404, 2024. 3
- [18] Zhi Huang, Federico Bianchi, Mert Yuksekogun, Thomas J Montine, and James Zou. A visual-language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 2, 3, 5
- [19] Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024. 2
- [20] Xinmi Huo, Kok Haur Ong, Kah Weng Lau, Laurent Gole, David M Young, Char Loo Tan, Xiaohui Zhu, Chongchong

- Zhang, Yonghui Zhang, Longjie Li, et al. A comprehensive ai model development framework for consistent gleason grading. *Communications Medicine*, 4(1):84, 2024. 6
- [21] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3
- [22] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology, 2023. 5
- [23] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 4
- [24] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018. 7
- [25] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023. 2
- [26] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3344–3354, 2023. 2
- [27] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo*10, 5281, 2018. 4, 6
- [28] Yoo Jung Kim, Hyungjoon Jang, Kyoungbun Lee, Seongkeun Park, Sung-Gyu Min, Choyeon Hong, Jeong Hwan Park, Kanggeun Lee, Jisoo Kim, Wonjae Hong, et al. Paip 2019: Liver cancer segmentation challenge. *Medical image analysis*, 67:101854, 2021. 6
- [29] Katharina Kriegsmann, Frithjof Lobers, Christiane Zgorzel-ski, Joerg Kriegsmann, Charlotte Janssen, Rolf Ruedinger Meliss, Thomas Muley, Ulrich Sack, Georg Steinbuss, and Mark Kriegsmann. Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology*, 12:1022967, 2022. 4
- [30] Vinay Kumar, Abul K Abbas, Nelson Fausto, and Jon C Aster. *Robbins and Cotran pathologic basis of disease, professional edition e-book*. Elsevier health sciences, 2014. 1
- [31] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021. 7
- [32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 5
- [33] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 4
- [34] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. 1, 2
- [35] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, pages 1–3, 2024. 1, 2
- [36] OpenAI. Gpt-4v(ision) system card. [https://cdn.openai.com/papers/GPTV\\_System\\_Card.pdf](https://cdn.openai.com/papers/GPTV_System_Card.pdf), 2023. 2, 5
- [37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2
- [38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 5
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2
- [40] Jeongun Ryu, Aaron Valero Puche, JaeWoong Shin, Seonwook Park, Biagio Brattoli, Jinhee Lee, Wonkyung Jung, Soo Ick Cho, Kyunghyun Paeng, Chan-Yong Ock, Donggeun Yoo, and Sérgio Pereira. Ocelot: Overlapped cell on tissue dataset for histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23902–23912, 2023. 3, 6
- [41] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. *arXiv preprint arXiv:2312.04746*, 2023. 2, 3
- [42] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192, 2024. 1
- [43] George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024. 2
- [44] Julio Silva-Rodríguez, Adrián Colomer, María A Sales, Rafael Molina, and Valery Naranjo. Going deeper through

- the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine*, 195: 105637, 2020. 4
- [45] Julio Silva-Rodríguez, Adrián Colomer, María A. Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine*, 195: 105637, 2020. 6
- [46] Dario Sitnik, Gorana Aralica, Mirko Hadžija, Marijana Popović Hadžija, Arijana Pačić, Marija Milković Periša, Luka Manojlović, Karolina Krstanac, Andrija Plavetić, and Ivica Kopriva. A dataset and a methodology for intraoperative computer-aided diagnosis of a metastatic colon cancer in a liver. *Biomedical Signal Processing and Control*, 66: 102402, 2021. 3, 6
- [47] Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng Lyu, Tao Lin, and Lin Yang. Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. *arXiv preprint arXiv:2407.00203*, 2024. 1, 2, 3, 5
- [48] Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5034–5042, 2024. 1, 2, 3, 5
- [49] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 5
- [50] Yuri Tolkach, Lisa Marie Wolgast, Alexander Damanakis, Alexey Pryalukhin, Simon Schallenberg, Wolfgang Hulla, Marie-Lisa Eich, Wolfgang Schroeder, Anirban Mukhopadhyay, Moritz Fuchs, et al. Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study. *The Lancet Digital Health*, 5(5):e265–e275, 2023. 3, 6
- [51] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018. 4
- [52] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023. 2
- [53] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings*, pages 11–24. Springer, 2021. 6
- [54] F. Wilm, M. Fragoso, C. Marzahl, C. Bertram, R. Klopffleisch, A. Maier, M. Aubreville, and K. Breininger. Canine cutaneous cancer histology dataset, 2022. 3, 6
- [55] Weidi Xie, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, and Yanfeng Wang. Towards generalist foundation model for radiology. 2023. 3
- [56] Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in Oncology*, page 4133, 2021. 3, 6
- [57] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10681–10690, 2019. 3, 6
- [58] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pages 1–8, 2024. 1, 2
- [59] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 2
- [60] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2, 4
- [61] Ying Zeng and Jialong Zhu. Transferring the knowledge of vision-language model for pathological image classification. In *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, pages 94–98, 2024. 6
- [62] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13, 2024. 3
- [63] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 5
- [64] Xiao Zhou, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pretraining for computational pathology. In *European Conference on Computer Vision*, pages 345–362. Springer, 2024. 5
- [65] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, Thomas Fuchs, Nicolo Fusi, et al. Virchow 2: Scaling self-supervised mixed magnification

models in pathology. *arXiv preprint arXiv:2408.00738*, 2024. [1](#), [2](#)

- [66] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition, 2018. [3](#), [6](#)