PathMMU: A Massive Multimodal Expert-Level Benchmark for Understanding and Reasoning in Pathology

Yuxuan Sun[1,2], Hao Wu[3], Chenglu Zhu[2], Sunyi Zheng[2], Qizi Chen[4], Kai Zhang[5], Yunlong Zhang[1,2], Dan Wan[6], Xiaoxiao Lan[1], Mengyue Zheng[2], Jingxiong Li[1,2], Xinheng Lyu[2], Tao Lin[2,†], and Lin Yang[2,†]

[1] Zhejiang University, Hangzhou 310058, China
[2] Westlake University, Hangzhou 310030, China
{sunyuxuan,zhuchenglu,lintao,yanglin}@westlake.edu.cn
[3] Macau University of Science and Technology, Macau 999078, China
2220024311@student.must.edu.mo
[4] Jiangnan University, Wuxi 214122, China
[5] The Ohio State University, Columbus 43210, USA
[6] Fujian University of Traditional Chinese Medicine, Fuzhou 350122, China
https://pathmmu-benchmark.github.io/

Abstract. The emergence of Large Multimodal Models (LMMs) has unlocked remarkable potential in AI, particularly in pathology. However, the lack of specialized, high-quality benchmark impeded their development and precise evaluation. To address this, we introduce PathMMU, the largest and highest-quality expert validated pathology benchmark for LMMs. It comprises 33,428 multimodal multi-choice questions and 24,067 images from various sources, each accompanied by an explanation for the correct answer. The construction of PathMMU leverages GPT-4V's advanced capabilities, utilizing over 30,000 image-caption pairs to enrich the descriptive quality of captions and generate corresponding Q&As in a cascading process. To maximize PathMMU's authority, we invite seven pathologists to scrutinize each question under strict standards in PathMMU's validation and test sets, while simultaneously setting an expert-level performance benchmark for PathMMU. We conduct extensive evaluations, including zero-shot assessments of 14 open-sourced and 4 closed-sourced LMMs and their robustness to image corruption. We also fine-tune representative LMMs to assess their adaptability to PathMMU. The empirical findings indicate that advanced LMMs struggle with the challenging PathMMU benchmark, with the top-performing LMM, GPT-4V, achieving only a 49.8% zero-shot performance, significantly lower than the 71.8% demonstrated by human pathologists. After fine-tuning, substantially smaller open-sourced LMMs can outperform GPT-4V but still fall short of the expertise shown by pathologists. We hope that the PathMMU will offer valuable insights and foster the development of more specialized, next-generation LMMs for pathology.

Keywords: Large multimodal model · Pathology · Benchmark

† Corresponding author.

1 Introduction

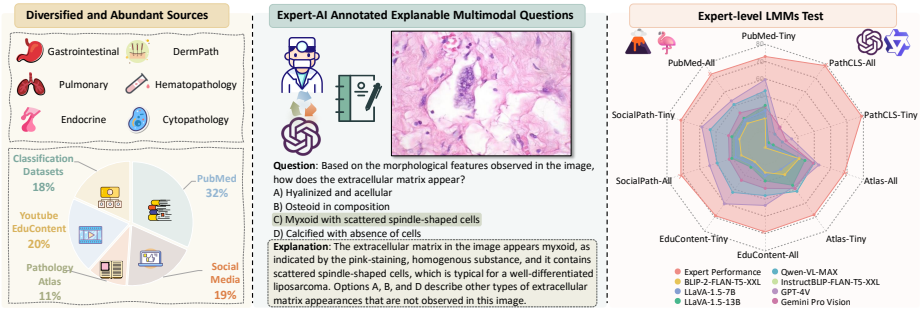


Fig. 1: An overview of the PathMMU benchmark: PathMMU is built from diverse, rich data sources. It comprises expert-level, multimodal, multiple-choice questions in pathology, collaboratively crafted by AI and pathology experts. The benchmark reveals even the most advanced LMMs significantly underperform compared to human experts.

Pathology is integral to modern medicine, serving as the foundation for diagnosing and understanding diseases. For instance, liquid-based cytology is crucial for early cervical cancer detection. Assessing biomarkers like HER2 in breast cancer guides the selection of targeted therapies and immunotherapies [23].

In recent years, the field of pathology has undergone a significant transformation, driven by advances in digital pathology and the integration of AI. This shift marks a departure from conventional microscope-based slide reading to AI-powered image analysis, greatly easing the workload of pathologists. Traditional pathology models are tailored for specific tasks, such as cervical cytology and liver lesion classification, leading to an abundance of task-specific models [57,59]. In contrast, recent advancements in Large Multimodal Models (LMMs) focus on offering general task-solving capabilities, thereby making universal pathological region recognition possible. This breakthrough not only represents a significant stride in the field, but also paves the way for more versatile and efficient diagnostic approaches in pathology [13, 27, 42, 44].

Given the demands of precise interpretation in pathology, conducting comprehensive evaluations of LMMs' abilities in interpreting pathology images is essential. However, the field faces a notable scarcity of high-quality benchmark datasets. Currently, the primary large-scale dataset available is PathVQA [18], which offers over 30,000 visual Q&A samples with 4,998 images. Nevertheless, these samples are derived from limited sources of textbooks, with image captions being converted into questions through heuristic approaches, constraining the generation of diverse and logical Q&As. Additionally, image captions in the textbook might not accurately reflect the corresponding images. For instance, some captions may introduce extraneous information not evident in the images or only partially cover the image content, thereby leading to unsolvable Q&As

or questions that inadequately capture the essence of the images. Moreover, the absence of expert review and validation during the data curation process may introduce substantial noise into the dataset. These challenges pose significant obstacles to effectively validating LMMs' capabilities in pathology.

To address these challenges, we introduce PathMMU, a multimodal expert-level benchmark designed to evaluate LMMs in pathology image understanding and reasoning, comprising 33,428 Q&As along with 24,067 pathology images. As depicted in Fig. 1, PathMMU draws from a diverse range of high-quality sources, including PubMed scientific papers, pathology atlas from guidelines, Twitter posts by pathology experts, commonly used pathology classification datasets and educational content from YouTube videos, covering multiple organ systems (*e.g.*, gastrointestinal, pulmonary, *etc.*) and multiple subjects (*e.g.*, dermatopathology, hematopathology, *etc.*). In developing PathMMU, we adopt a meticulously designed cascading approach. Initially, we prompt GPT-4V with collected image-caption pairs to enhance the original captions, crafting descriptions that delve deeper into details and highlight crucial morphological features. Subsequently, we use these enhanced descriptions along with the images to prompt GPT-4V to generate professional multi-choice, multimodal pathology Q&As, each accompanied by detailed explanations for their answers. To ensure the questions are specifically designed for multimodal pathology understanding, we employ a collaboration of multiple Large Language Models (LLMs) to eliminate questions that can be solved or guessed using text alone. Most importantly, we invite seven pathology experts to manually review approximately 12,000 Q&As from the test and validation sets of PathMMU. The above carefully executed procedures, ensure the generation of professional, logical, and high-quality Q&As.

We evaluate the zero-shot performance of 18 advanced LMMs and their robustness on the PathMMU test sets. We also fine-tune two representative LMMs on the training set to assess their transfer learning capabilities. The key insights from this extensive analysis are: **(1) Cutting-edge LMMs struggle with the PathMMU**, with 15 out of the 18 achieving no more than 40% accuracy. Even the most advanced model currently available, GPT-4V, only attains an accuracy of 49.8%. This reveals a notable discrepancy of nearly 20% compared to professional pathologists, indicating significant deficiencies of current LMMs in the specialized field of pathology and emphasizing the challenging nature of the PathMMU benchmark. **(2) While these LMMs show considerable robustness in robustness tests, the full extent of their robustness is still in question.** On the one hand, given their limited performance, the scope for further performance degradation is somewhat constrained. On the other hand, these models tend to capture only superficial image features, and sometimes may not even utilize image information for reasoning. As a result, even when image corruption obscures minor details, the impact on these models' performance might not be significantly pronounced. **(3) The text-only performance of GPT-4 Turbo surpasses the top-performing open-source LMMs on the PathMMU benchmark. We observe that GPT-4 Turbo sometimes takes shortcuts to guess answers.** This involves choosing the correct answer based

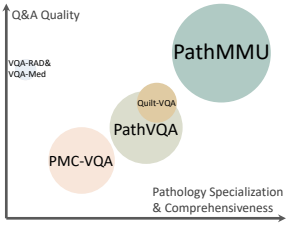
either on the options most commonly encountered in real-world pathological scenarios, or by identifying the most distinctive option, rather than conducting an in-depth analysis of the pathology image. **(4) Fine-tuning general-purpose LMMs with substantial amounts of data significantly boosts their ability to comprehend and reason with pathological images**, enabling them to easily surpass the zero-shot performance of GPT-4V. However, there still remains a discernible gap between their performance and that of human experts.

2 Related Works

Multimodal Models. The emergence of powerful large language models such as BERT [12], GPT-3 [8], T5 [38], LLaMa [46], ChatGPT [33], GPT-4 [34], and Vicuna [10] has significantly advanced the field of Natural Language Processing (NLP). These models demonstrate exceptional general capabilities, motivating researchers to explore the integration of LLMs with vision models to develop versatile multimodal models. This integration is primarily achieved through the pretraining approach, leading to the creation of groundbreaking LMMs. Representative models of this approach include CLIP [37], Flamingo [1], BLIP-2 [29], and Fuyu [5], all of which demonstrate remarkable multimodal understanding abilities. Furthermore, by adapting instruction-tuning techniques from the NLP domain, a variety of multimodal instruction-tuning datasets have been developed. This initiative aims to guide LMMs in generating outputs that are more controllable, practical, and adaptable to various tasks. Simultaneously, this approach facilitates the integration of LLMs into LMMs using lightweight fine-tuning, significantly reducing both costs and time. Key models that embody this advancement include GPT-4V [35], Gemini Pro Vision [45], Qwen-VL [4], LLaVA [30], MiniGPT-4 [60], BLIP-2 [29] and InstructBLIP [11].

LMM Benchmarks. In the general domain, the rise of various large models has led to the creation of extensive benchmark datasets, designed to evaluate the universal capabilities of these models across a range of tasks and domains. Notable examples include LAMM [52], LVLM-eHub [51], SEED [26], MMBench [31], MM-Vet [53] and BenchLMM [9], which have been used to assess the basic perception abilities of large models. More recently, MMMU [54], a massive and challenging dataset covering 30 university-level subjects, has been developed. It is specifically designed to evaluate the general multimodal understanding and reasoning capabilities of LMMs. Furthermore, HaELM [48] and HallusionBench [16] are proposed to evaluate the degree of hallucination in LMMs.

In the medical field, datasets such as VQA-RAD [24] and VQA-Med [6] provide a collection of Q&A pairs based on radiology images, facilitating the evaluation of the capabilities of AI in radiological diagnostics. Additionally, PMC-VQA [55] is a large-scale dataset within this domain. It generates a vast number of Q&A pairs by prompting ChatGPT with text-only image captions from PubMed. In the pathology domain, limited dataset construction efforts have been performed to date. The representative large-scale dataset available, PathVQA [18], is constructed in a manner similar to PMC-VQA. It utilizes



Dataset	Domain	Sources	# Q&As / Images	Answer Explanation?	LLM Filtered?	Expert Annotated?
VQA-Med	Medical	MedPix	5000 / 5000	✗	✗	✓
VQA-RAD	Medical	MedPix	3515 / 315	✗	✗	✓
PMC-VQA	Medical	PMC	227k / 149k	✗	✗	✗
PathVQA	Pathology	Book	32799 / 4998	✗	✗	✗
Quilt-VQA Quilt-Red	Pathology	Edu-Content	1283 / 985	✗	✗	✓
PathMMU	Pathology	Social Media Edu-Content, PubMed Atlas, CLS-Data	33428 / 24067	✓	✓	✓

Fig. 2: The comparison between PathMMU and existing benchmarks. The Q&A pairs in PathMMU are sourced extensively and comprehensively, undergoing rigorous multi-tiered filtering. This includes the initial filtering by multiple LLMs and the strict reviews by professional pathologists. Additionally, each question is accompanied by a detailed explanation. These attributes establish PathMMU as the most professionally curated, comprehensive, and highest-quality large-scale pathology dataset available.

a heuristic approach to generate questions based on the text-only captions of pathology textbook images, yielding over 30,000 Q&As and 4,998 images. More recently, Quilt-VQA [39] is proposed, which is constructed by automatically extracting question and answer pairs from the narrations of professional teachers in YouTube videos. This approach guarantees the professionalism of the Q&A pairs, although it comprises a relatively smaller dataset with 1,283 Q&As and 985 images. Nonetheless, datasets reliant solely on image captions for question generation, face several notable limitations: **(1)** The creation of questions typically requires the consideration of both the image and its accompanying caption. Relying only on captions can introduce inaccuracies, making some questions unanswerable or solvable using only the text, bypassing the need for image analysis. **(2)** Since captions tend to be simplistic and may not capture the detailed information evident in images, generating an excessive number of Q&As from the same image's caption can lead to either too simplistic or insufficient questions. **(3)** Automatically generated questions, whether through predefined rules or by prompting ChatGPT, are susceptible to containing numerous errors.

Given the current lack of high-quality benchmarks in pathology, we develop PathMMU, a large-scale, high-quality, and expert-verified pathology benchmark, to fill this gap. Our approach involves generating Q&As from a diverse range of image-caption pairs, followed by rigorous expert review and filtering applied to both the images and the questions. Moreover, we provide detailed explanations for each answer as a reference to enhance the interpretability of answers, which is absent in previous works. As shown in Fig. 2, PathMMU distinguishes itself in the pathology domain through its professionalism and superior quality.

3 The Proposed PathMMU Benchmark

3.1 Design Principle

PathMMU is designed to provide the community with a specialized dataset for evaluating pathology LMMs. We adhere to five principal guidelines in its devel-

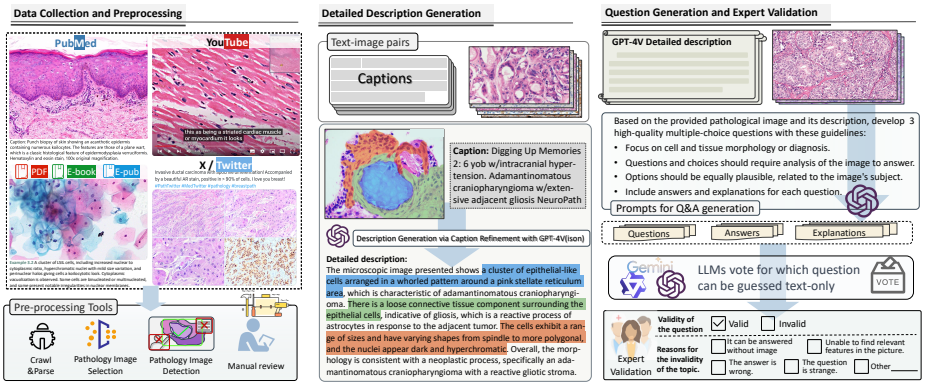


Fig. 3: An illustrative overview of the three main processes in PathMMU Q&A generation: data collection and preprocessing, detailed pathology image description generation, and question generation with LLMs filtering and expert validation.

opment: **(1) Comprehensive and Specialized Data**: Our benchmark is gathered from a diverse range of authoritative materials, including respected scientific publications from PubMed, pathology atlas from textbooks and guidelines, educational videos on YouTube, pathologist-shared images with explanations on Twitter, and widely recognized pathology classification datasets. Additionally, we involve seven human experts to manually review the dataset to ensure it meets professional standards. **(2) Effective and Valuable Questions**: The questions in PathMMU are carefully formulated to necessitate answers based on detailed observation of pathology images, rather than being solvable through mere textual interpretation or guesswork. We ensure that these questions are answerable, effective, and aligned with the standards of professional pathology examinations. **(3) Large-Scale**: We present the largest pathology dataset currently available. This scale enables researchers to thoroughly explore the full potential of LMMs in the field of pathology. **(4) High-Quality Images**: We prioritize clarity in our images to ensure that every detail is discernible, as evidenced by our dataset's average resolution of approximately 900 $\times$ 700 pixels. **(5) Explainable Answer Choices**: In line with the ongoing interest in understanding the decision-making logic of large models, our dataset includes reference explanations for each answer choice. This feature enhances the interpretability of model responses, contributing to the broader research on model explainability.

3.2 The Construction of PathMMU

To guarantee the quality of our benchmark, we meticulously develop a three-step data processing and generation protocol. The overall data collection and generation process is illustrated in Fig. 3.

Step 1: Data Collection and Preprocessing. Our data collection draws from a wide variety of sources, which we integrate as subsets into the PathMMU framework. We name these subsets and detail their collection workflows as follows.

PubMed: This subset consists of scientific pathology image-text pairs sourced from scientific documents. We gather these pairs from the open-access section of PubMed Central. Given the significant presence of non-pathological data within this resource, we meticulously annotate 20,000 samples as either pathological or non-pathological, and subsequently train a ConvNeXt [32] model to identify pathology data within the remaining dataset. **EduContent**: This subset, derived from educational YouTube videos, is sourced from Quilt1M [20]. Since YouTube videos often include a variety of visuals unrelated to pathology, such as computer desktops and unrelated imagery, the initial phase involves the annotation of pathological regions in 3,000 YouTube video images. This critical step is followed by the training of a YOLOv6-based detector [28] to automate the identification of pathological regions within remaining YouTube content. **Atlas**: This subset is a compilation of authoritative pathology textbooks and guidelines, we transform them from PDFs to HTML to facilitate the extraction of image-caption pairs. **SocialPath**: This subset aggregates image-text pairs from Twitter posts by pathology experts, using Twitter URLs provided by OpenPath [19]. **PathCLS**: Encompassing widely recognized pathology classification datasets, including PatchCamelyon [47], CRC-100K [21], SICAPv2 [40], BACH [2], Osteo [3], SkinCancer [22], MHIST [50], WSSSLUAD [17], LC-Lung and LC-Colon [7]. We reformulate these datasets into a Q&A format and randomly sample from these datasets. We directly make the gathered samples from PathCLS a part of the PathMMU dataset without undergoing the following two steps Q&A generation process. Finally, we manually review and filter irrelevant or unclear images to ensure the quality of collected data. As a result, approximately 30,000 high-quality image-text pairs are obtained, forming the foundation of the PathMMU.

Step 2: Detailed Description Generation and Question-answer Pairs Generation. When dealing with data sources from platforms like Twitter and YouTube, it's common to find a weak correlation between an image and its accompanying caption. For instance, in the SocialPath dataset, captions might only describe the aesthetic appeal of the image, which is irrelevant to pathological findings. Similarly, in other data sources, captions tend to describe only a fraction of the features present in the image. To tackle this issue, we prompt GPT-4V to focus on describing the morphology of cells and tissues. It is crucial to note that merely prompting GPT-4V to output image descriptions might lead to incoherent content. Therefore, we provide GPT-4V with original captions for reference, significantly reducing the incidence of such occurrences and thus ensuring the generation of more precise and relevant descriptions.

Step 3: Question Generation and Expert Validation. Following the generation of image descriptions, GPT-4V is tasked with creating three questions per image, each accompanied by multiple-choice options, the correct answer, and detailed explanations for the answer. The creation of these explanations serves a dual purpose: it not only offers valuable insights into the model reasoning process, but also facilitates the subsequent manual review phase.

It's crucial to note that despite the careful design of question generation prompts for GPT-4V, the text-only GPT-4 is capable of correctly answering

over 60% of these questions through educated guesses, bypassing the need for visual cues. GPT-4 explains that its deductions are drawn from patterns such as one option being more common in typical pathological scenarios or exhibiting noticeable differences from other options. We delve deeper into this phenomenon in Sec. 4.3 to offer a more thorough experimental analysis. To ensure that LMMs do not rely on educated guesses, thereby obscuring their true multimodal capabilities, we employ GPT-3.5 Turbo, GPT-4 Turbo, Gemini Pro, and ERNIE-Bot-4 to perform educated guesses. Questions correctly guessed by at least three of these models are subsequently excluded.

We partitioned the remaining dataset into training, validation, and test sets. The training set comprises 16,344 images and 23,041 Q&As, while the validation and test sets contain roughly 12,000 Q&As and over 8,000 images, respectively. For the validation and test sets, we invite seven professional pathologists to conduct a thorough manual review. These pathologists initially attempt to answer the questions independently and then assess them based on the following criteria: **(1)** Whether the question can be answered without an accompanying image. **(2)** Whether the answer can be inferred from the provided question and image. **(3)** Whether the supplied answer is incorrect, if there is no correct answer, or if multiple correct answers are possible. **(4)** Whether the generated question appears unusual or atypical compared to standard questions in pathology examinations. Questions failing to meet these standards are deemed invalid and consequently removed from the PathMMU dataset.

Ultimately, after thorough expert review, we obtain 710 Q&As accompanied by 510 images for the validation set, and 9,677 Q&As with 7,213 images for the test set. Additionally, to establish a standard for expert performance, we extract a smaller subset from the test set, named 'test-tiny', which includes 1,156 Q&As. We invite two groups of professional pathologists to participate in this subset as an examination, and we average their performance to set a benchmark for expert performance in the PathMMU, serving as a comparison reference for LMMs.

4 Experiments

In this section, we conduct extensive experiments on PathMMU. Initially, we assess the zero-shot performance of cutting-edge LMMs. To verify whether LMMs effectively utilize the visual information from pathology images, we also evaluate the performance of text-only LLMs on PathMMU as a reference. Additionally, we apply common corruptions to the images in the test-tiny set to test the robustness of the LMMs against image corruptions. Furthermore, we select representative LMMs and evaluate their performance to explore their transfer-learning capabilities. Finally, we conduct experimental analysis on the issue of LLMs being able to guess answers identified during the data construction process.

4.1 Zero-shot Evaluation of LMMs and LLMs

In this study, we evaluate the zero-shot capabilities of the latest and most advanced LMMs on the PathMMU. Specifically, we use the PathMMU validation

Table 1: Overall results of models on the PathMMU **validation** and **test set**. Besides reporting the performance of LMMs, we add text-only LLM baselines that purely accept text as inputs. The best-performing LMM in each subset is **in-bold**, and the top-performing LLM is underlined.

	Validation Overall	Test Overall		PubMed		SocialPath		EduContent		Atlas		PathCLS	
	- (710)	Tiny (1156)	ALL (9677)	Tiny (281)	ALL (3068)	Tiny (235)	All (1855)	Tiny (255)	All (1938)	Tiny (208)	ALL (1007)	Tiny (177)	ALL (1809)
Random Choice	24.6	22.1	23.7	22.1	25.1	25.5	26.5	25.5	26.0	19.7	23.0	15.3	16.3
Frequent Choice	27.5	27.7	25.5	28.8	26.1	27.7	26.7	29.8	26.5	28.4	27.5	22.0	21.0
Expert performance	-	71.8	-	72.9	-	71.5	-	69.0	-	68.3	-	78.9	-
Large Multimodal Models (LMMs): Text + Image as Input													
OpenFlamingo2-9B [1]	26.2	24.1	24.2	24.9	25.5	25.5	25.3	19.6	24.4	28.8	25.5	22.0	20.4
Kosmos2 [36]	26.9	26.1	24.9	27.8	25.8	27.2	24.3	25.1	25.9	23.1	24.2	27.1	22.8
LLaMA-Adapter2-7B [14]	26.2	26.6	26.4	26.0	28.3	27.7	26.5	27.1	28.0	30.8	27.4	20.9	20.5
MiniGPT4-Vicuna-13B [60]	27.2	25.5	27.7	28.8	30.1	24.3	27.2	25.5	29.3	27.9	28.6	19.2	21.8
Otter [25]	26.1	29.3	28.1	34.9	30.2	26.0	28.4	30.6	30.0	29.3	28.2	23.2	21.3
Fuyu-8B [5]	27.5	30.1	29.2	35.9	30.5	28.9	30.7	30.6	29.8	27.9	30.6	24.3	23.9
CogVLM [49]	29.9	30.6	29.7	32.0	32.1	31.1	30.4	30.6	29.9	35.1	33.7	22.6	22.4
Qwen-VL-7B [4]	29.4	32.1	31.5	34.9	32.9	33.6	35.8	34.1	33.6	35.6	33.8	18.6	21.7
BLIP-2 FLAN-T5-XL [29]	32.8	32.6	31.8	35.9	34.6	34.9	33.6	33.7	32.8	34.6	37.4	20.3	20.8
BLIP-2 FLAN-T5-XXL [29]	33.4	33.3	33.5	37.0	37.4	35.7	34.6	30.2	34.5	39.4	40.7	19.8	20.6
InstructBLIP FLAN-T5-XL [11]	33.1	34.9	31.8	37.0	33.2	35.3	33.9	36.5	33.3	37.0	36.7	26.6	23.1
InstructBLIP FLAN-T5-XXL [11]	32.7	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-1.5-7B [30]	36.6	34.9	35.4	41.6	39.9	37.9	38.1	32.5	36.5	35.6	39.2	23.2	21.9
LLaVA-1.5-13B [30]	37.9	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-PLUS [4]	38.0	39.3	34.3	43.5	37.7	41.3	36.0	39.6	36.0	44.7	37.1	23.2	23.3
Qwen-VL-MAX [4]	43.6	49.2	45.9	53.0	50.9	53.6	49.3	52.2	47.9	**51.4**	49.8	30.5	29.6
Gemini Pro Vision [45]	41.9	42.8	42.7	43.8	44.9	42.4	42.0	43.5	43.7	49.5	49.4	32.8	**34.7**
GPT-4V-1106 [35]	**49.3**	**53.9**	**49.8**	**59.4**	**53.5**	**58.7**	**53.9**	**60.4**	**53.6**	48.1	**52.8**	**36.2**	33.8
Large Language Models (LLMs): Only Text as Input													
ERNIE-Bot 4.0 [41]	31.8	34.3	30.9	37.0	31.2	33.6	32.9	40.0	34.5	36.1	37.4	20.9	20.6
Gemini Pro [45]	31.0	31.6	31.0	31.3	33.5	31.9	31.4	33.3	32.7	38.0	37.7	21.5	20.9
Vicuna-v1.5-13B [10]	32.0	31.2	31.6	35.2	33.8	30.6	34.3	31.0	33.5	34.6	35.7	22.0	20.7
GPT-3.5 Turbo [33]	30.4	29.4	28.4	32.7	31.2	31.5	30.1	31.0	30.1	26.9	26.8	22.0	20.8
GPT-4 Turbo [34]	36.5	41.8	38.1	48.8	42.4	43.4	42.3	47.1	40.6	41.3	43.2	22.3	21.1

set to conduct prompt engineering for these models, and then test their performance on the test set. The tested models include 14 open-source LMMs: OpenFlamingo [1], Kosmos2 [36], LLaMA-Adapter2 [14], MiniGPT-4 [60], Otter [25], Fuyu [5], CogVLM [49], Qwen-VL [4], BLIP-2 [29], InstructBLIP [11], and LLaVA-1.5 [30], as well as 4 closed-source LMMs: Qwen-VL-PLUS, Qwen-VL-MAX [4], Gemini Pro Vision [45], and GPT-4V [35]. For BLIP-2, InstructBLIP, and LLaVA-1.5, we deploy various sizes of these models to examine the applicability of scaling rules in LMMs. Moreover, we extend our evaluation to purely text-based LLMs, including ERNIE-Bot 4.0 [41], Gemini Pro [45], Vicuna-v1.5-13B [10], GPT-3.5 Turbo [33], and GPT-4 Turbo [34]. This assessment aims to quantify their ability to infer correct answers through educated guesses without visual inputs, reflecting the extent to which LMMs integrate and leverage visual information compared to their text-only counterparts.

Results and Discussion. ***The advanced LMMs struggle with the PathMMU dataset***. As shown in Tab. 1, among the 18 models tested, 15 show an accuracy below 40%. Notably, the highest-performing open-source model, LLaVa-1.5-13B, and the top-performing closed-source model, GPT-4V, achieve

only 37.6% and 49.8% accuracy, respectively. Those are significantly lower than human expert performance, which stands at 71.8%, underscoring a substantial gap between current LMMs and professional pathologist standards. This gap indicates that the practical application of LMMs in real-world pathological scenarios remains significantly constrained.

Closed-source LLMs achieve performance that is on par with or even surpasses that of the most advanced open-source LMMs. When considering text-only models, we prompt them to make educated guesses without providing the image. We observe that GPT-4 Turbo, Vicuna-v1.5-13B, Gemini-Pro, and ERNIE-Bot 4.0 outperform nearly half of the open-source LMMs. Notably, GPT-4 Turbo leads with a significant margin, achieving an accuracy rate of 38.1%, exceeding the second-best LLM by 6.5%, and even outperforming the best-performing open-source LMM, LLaVA-1.5-13B. The 38.1% accuracy of GPT-4 Turbo, on the one hand, demonstrates that under our strict filtering criteria, even the best LLM, when provided with text alone, makes incorrect answers for the majority of questions. This underscores the effectiveness of our dataset in testing the capabilities of LMMs in pathological image analysis. On the other hand, it reaffirms the analysis in Sec. 3.2 that GPT-4 possesses strong logical reasoning abilities, enabling it to correctly guess a portion of the questions, significantly outperforming random selection. However, this raises critical concerns in pathology diagnosis: the reliability of model outputs in clinical scenarios remains questionable, especially when they rely on educated guesses, particularly probabilistic ones, rather than comprehensive analysis. This emphasizes the urgent necessity to develop interpretable AI models and advocate for their cautious application in pathology diagnostics.

The larger LMMs exhibit better performance. Specifically, LLaVa-1.5-13B exceeds LLaVa-1.5-7B model by 2.2% in overall performance, while InstructBLIP FLAN-T5-XXL and BLIP-2 FLAN-T5 XXL surpass their smaller versions by 2.1% and 1.7%, respectively. This suggests that larger models typically possess stronger multimodal capabilities in the field of pathology.

4.2 Robustness of LMMs to Image Corruption

In practical pathology, the interpretation of models significantly influences the subsequent medical decisions and treatment strategies. Therefore, models with strong robustness are crucial for clinical applications. The quality of pathological slides can be affected by various factors during staining, scanning, and storage, including JPEG compression, pixelation, blur (*e.g.*, bubble blur, defocus blur, and motion blur), and color variations (*e.g.*, brightness, saturation, and hue).

Drawing inspiration from the studies on the robustness of encoder-based models to pathology image corruptions [43, 56, 58], we incorporate these aforementioned types of corruptions into our analysis. To be specific, we simulate five levels of each corruption type on the PathMMU to explore the robustness of LMMs against these corruptions, as depicted in the left half of Fig. 4.

Results and Discussion. ***LMMs demonstrate notable robustness to various types and levels of image corruptions, yet their true robustness***

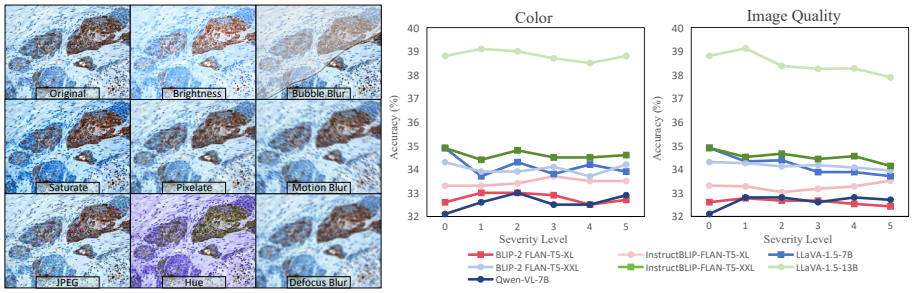


Fig. 4: Left: Illustration of corrupted pathology images. Right: LMM's performance across various levels of color-related (brightness, hue, saturation) and image quality-related (pixelation, JPEG compression, bubble blur, motion blur, defocus blur) corruptions on the PathMMU test-tiny set, with level 0 representing the uncorrupted images.

Table 2: Evaluation of the model's robustness on the PathMMU test-tiny set, where green indicates performance improvement and red signifies performance degradation compared to the model's performance with uncorrupted images.

	Brightness	Bubble	Defocus	Hue	JPEG	Motion	Pixelate	Saturate	Overall
Qwen-VL-7B	32.7 ↑ 0.6	32.3 ↑ 0.2	32.1 ↓ 0.0	32.4 ↓ 0.3	32.3 ↑ 0.2	33.8 ↑ 1.7	33.1 ↑ 1.0	33.0 ↑ 0.9	32.7 ↑ 0.6
BLIP-2 FLAN-T5-XL	32.9 ↑ 0.3	32.4 ↓ 0.2	32.8 ↑ 0.2	32.9 ↑ 0.3	32.4 ↓ 0.2	32.8 ↑ 0.2	32.6 ↑ 0.0	32.7 ↑ 0.1	32.7 ↓ 0.1
BLIP-2 FLAN-T5-XXL	33.1 ↓ 0.2	33.2 ↓ 0.1	33.5 ↑ 0.2	33.7 ↑ 0.4	33.5 ↑ 0.2	33.1 ↓ 0.2	33.1 ↓ 0.2	33.6 ↑ 0.3	33.3 ↓ 0.0
InstructBLIP FLAN-T5-XL	34.0 ↓ 0.9	34.1 ↓ 0.8	33.9 ↓ 1.0	34.1 ↓ 0.8	33.7 ↓ 1.2	34.0 ↓ 0.9	34.4 ↓ 0.5	33.8 ↓ 1.1	34.0 ↓ 0.9
InstructBLIP FLAN-T5-XXL	34.0 ↓ 0.3	33.8 ↓ 0.5	34.2 ↓ 0.1	34.1 ↓ 0.2	34.3 ↓ 0.2	34.3 ↑ 0.0	34.1 ↓ 0.2	33.8 ↓ 0.5	34.1 ↓ 0.2
LLaVA-1.5-7B	34.8 ↓ 0.1	35.3 ↑ 0.4	34.1 ↓ 0.8	34.3 ↓ 0.6	34.0 ↓ 0.9	34.3 ↓ 0.6	34.6 ↓ 0.3	34.7 ↓ 0.2	34.5 ↓ 0.4
LLaVA-1.5-13B	38.9 ↑ 0.1	38.3 ↓ 0.5	38.3 ↓ 0.5	38.4 ↓ 0.4	38.0 ↓ 0.8	38.6 ↓ 0.2	38.7 ↓ 0.1	39.1 ↑ 0.3	38.5 ↓ 0.3

Table 3: Results of LMMs on the PathMMU test set with the original images substituted by random Gaussian noise images.

	Overall		PubMed		SocialPath		EduContent		Atlas		PathCLS	
	Tiny	ALL	Tiny	ALL	Tiny	All	Tiny	All	Tiny	ALL	Tiny	ALL
Qwen-VL-7B	28.8 ↓ 3.3	28.3 ↓ 3.2	39.5 ↑ 4.6	30.4 ↓ 2.5	24.7 ↓ 8.9	30.5 ↓ 5.3	27.8 ↓ 6.3	28.9 ↓ 4.7	27.4 ↓ 8.2	30.1 ↓ 3.7	20.3 ↑ 1.7	20.8 ↓ 0.9
BLIP-FLAN-T5-XL	30.9 ↓ 1.7	30.8 ↓ 1.0	30.6 ↓ 5.3	33.1 ↓ 1.5	34.5 ↓ 0.4	33.4 ↓ 0.2	32.2 ↓ 1.5	31.9 ↓ 0.9	34.1 ↓ 0.5	35.1 ↓ 2.3	20.9 ↑ 0.6	20.8 ↓ 0.0
BLIP-FLAN-T5-XXL	32.4 ↓ 0.9	31.6 ↓ 1.9	34.5 ↓ 2.5	34.6 ↓ 2.8	34.0 ↓ 1.7	32.5 ↓ 2.1	30.6 ↑ 0.4	32.8 ↓ 1.7	38.5 ↓ 0.9	38.1 ↓ 2.6	22.2 ↑ 2.4	20.8 ↑ 0.2
InstructBLIP FLAN-T5-XL	31.1 ↓ 3.8	29.7 ↓ 2.1	32.4 ↓ 4.6	30.8 ↓ 2.4	32.8 ↓ 2.5	32.0 ↓ 1.9	33.7 ↓ 2.8	32.4 ↓ 0.9	32.7 ↓ 4.3	30.4 ↓ 6.3	21.5 ↓ 5.1	22.4 ↓ 0.7
InstructBLIP FLAN-T5-XXL	30.4 ↓ 3.9	30.8 ↓ 3.1	32.0 ↓ 7.1	33.5 ↓ 3.7	28.9 ↓ 4.7	31.9 ↓ 2.4	32.2 ↓ 2.3	31.7 ↓ 4.3	35.1 ↓ 3.4	35.2 ↓ 4.1	21.5 ↓ 1.1	21.7 ↓ 1.0
LLaVA-1.5-7B	30.2 ↓ 4.7	30.9 ↓ 4.5	37.0 ↓ 4.6	33.9 ↓ 6.0	31.5 ↓ 6.4	32.5 ↓ 5.6	28.2 ↓ 4.3	32.7 ↓ 3.8	29.8 ↓ 5.8	33.9 ↓ 5.3	20.9 ↓ 2.3	20.8 ↓ 1.1
LLaVA-1.5-13B	32.7 ↓ 6.1	33.2 ↓ 4.4	38.4 ↓ 6.1	36.0 ↓ 5.0	35.3 ↓ 5.1	36.3 ↓ 4.1	29.0 ↓ 5.1	34.1 ↓ 5.3	35.6 ↓ 11.5	38.8 ↓ 5.5	22.0 ↓ 2.9	21.0 ↓ 2.5

is questionable, as shown in Tab. 2 and the right half of Fig. 4. Notably, the Qwen-VL-7B even shows a 0.6% overall performance increase under corruption compared to its baseline. It is more plausible to hypothesize that these corruptions mainly alter minute details of the pathology image that are difficult for LMMs to discern (such as chromatin morphology of the nucleus or vacuolization of the cytoplasm). This limitation stems from their pre-training in general domains, where they learn to recognize larger and more prominent features (*e.g.*, humans, houses, cars, *etc.*) rather than the nuanced details that are vital in pathology. Furthermore, they may even resort to exploiting spurious correlations as shortcuts to answer questions (*e.g.*, deducing answers from textual patterns instead of focusing on essential image details.)

To substantiate our argument, we employ an extreme test of corruption by replacing the images with Gaussian noise-generated random images. As shown in Tab. 3, we discover that even when the images contain no relevant information, LMMs can still achieve results significantly better than random choice. The drop in performance when using these random images ranged from only 1.0% to 4.5%. This finding suggests that LMMs might rely on shortcuts to accomplish the multimodal task, such as depending exclusively on textual information to make predictions. Additionally, it is intriguing that the performance drop across different sizes of the same model is remarkably similar. In other words, the benefit that images provide to models of various model sizes appears to be consistent, suggesting that the improvement in performance for different-sized models may primarily stem from their LLM component, rather than the vision aspect.

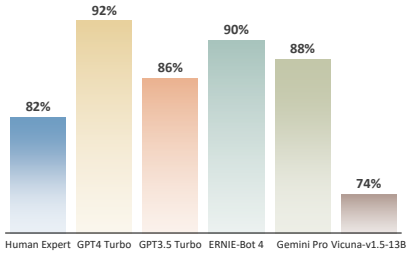
4.3 Analysis of LLMs' Ability to Guess Answers

In this section, we delve deeper into the phenomenon of LLM's ability to make educated guesses with more comprehensive experiments and analysis.

To more clearly demonstrate the guessing abilities of LLMs, we randomly select 100 samples that are filtered during the Q&A generation process and can be correctly guessed by multiple LLMs. We invite pathology experts to answer these questions with reference images. As shown in Fig. 5, despite having access to images, their performance is significantly lower than that of the closed-source LLMs, which achieve about 90% accuracy, and only slightly better than the smaller open-source Vicuna-v1.5-13B. This indicates that LLMs may outperform humans in guessing correct answers by identifying shortcuts within the questions.

We hypothesize several key reasons for the guessing behavior exhibited by LLMs: ***(1) Frequency-based guessing:*** LLMs may guess based on the prevalence of certain options in real-world pathological instances. ***(2) Based on the options that present a pattern of one supporting and three contradicting, or vice versa:*** For example, when a question identifies a tumour diagnosis, one option may match the pathological traits of a lesion, while the others suggest non-lesional characteristics. ***(3) When only one option aligns with the question's subject:*** For example, if a question asks which feature in the image supports the diagnosis of a Low-grade Squamous Intraepithelial Lesion (LSIL) cell, with only one option describes LSIL's pathological characteristics.

To support our hypothesis, we expand the sample size to 1,600, with 400 from each source (except PathCLS), to empirically analyze how LLMs guess answers. Specifically, we swap the questions among these samples while keeping the options unchanged, creating samples where the questions and options are completely mismatched. As shown in the right half of Fig. 5, we observe that LLMs still manage to guess approximately 50% cases correctly, significantly higher than random choice accuracy. This finding suggests that the models tend to select the most common or prominent option as the answer, supporting our hypotheses (1) and (2). To further explore hypothesis (3), we design experiments using BERT-large [12] and BiomedBERT-large [15], the latter being specifically pre-trained on biomedical data. Given that the BERT series incorporates Next-Sentence



	PubMed	SocialPath	EduContent	Atlas	Overall
Random Choice	25.8	24.3	25.5	25.8	25.3
GPT-4 Turbo	93.0	91.8	86.8	83.8	88.8
+ Random Swap Question	44.3	45.5	50.0	51.5	47.8
GPT-3.5 Turbo	84.3	86.5	79.5	85.3	83.9
+ Random Swap Question	44.0	44.5	53.8	51.3	48.4
Gemini Pro	85.5	90.1	76.5	71.5	84.1
+ Random Swap Question	43.8	44.8	50.3	51.3	47.5
ERNIE-Bot 4.0	94.5	92.3	89.3	82.0	89.5
+ Random Swap Question	47.0	46.5	56.5	51.0	50.3
BERT-large w/ NSP	24.5	36.0	37.5	34.3	33.1
BiomedBERT-large w/ NSP	44.3	44.8	42.3	37.5	42.3

Fig. 5: Left: The performance comparison between different LLMs and human experts on 100 filtered samples where the answer can be guessed through text-only. Right: Expand the sample size to 1600 to validate the source of LLMs' ability to guess answers, which includes: (1) randomly replacing the original questions with others from the dataset while keeping the options unchanged, and (2) utilizing the BERT series for answer selection, specifically through its Next Sentence Prediction (NSP), to assess whether an option is the sequential sentence following a question.

Prediction (NSP) in their pretraining, which inherently relies on the similarity between two sentences, we apply this to predict the relationship between questions and options. We select the option with the highest probability of being the correct next sentence to the question. Our findings reveal that both BERT and BiomedBERT substantially surpass random guessing, indicating that direct matching of questions and options is a viable method for models to guess the correct answers, thereby supporting hypothesis (3). Moreover, BiomedBERT-large demonstrates a notable performance enhancement compared to BERT-large. This suggests that pretraining on biomedical data equips the model with a broader understanding of pathological knowledge.

4.4 Fine-tuning Results

To explore the adaptability of LMMs to the pathology domain, we select representative LMMs, InstructBLIP FLAN-T5-XL and InstructBLIP FLAN-T5-XXL for fine-tuning experiments. Our experiments consist of two parts: (1) training the LMMs to directly generate answers, and (2) fine-tuning the LMMs to generate a reasoning process before delivering the final answer.

Results and Discussion. ***All models exhibit significant improvements on the PathMMU test set after fine-tuning on its training set***, as detailed in Tab. 4. Notably, during the fine-tuning for direct answer generation, InstructBLIP FLAN-T5-XL and InstructBLIP FLAN-T5-XXL achieve significant improvements of 21.5% and 21.3%, respectively. This substantial improvement enables them to outperform the current leading model, GPT-4V. While still lagging behind expert performance, these results demonstrate a trend toward approaching expert-level proficiency. This also reflects the effectiveness of PathMMU in enhancing the LMMs' abilities for pathology image analysis.

Unexpectedly, generating explanations before answers during fine-tuning does not yield improvements. Instead, we observe slight performance decreases of 0.3%

Table 4: Results of LMMs on the PathMMU test set after the fine-tuning. The 'w/ A' and 'w/ A&E' denote the model is fine-tuned to output the answer directly or to output the answer with an explanation for its answer.

	Overall		PubMed		SocialPath		EduContent		Atlas		PathCLS	
	Tiny	ALL	Tiny	ALL	Tiny	All	Tiny	All	Tiny	ALL	Tiny	ALL
InstructBLIP FLAN-T5-XL	34.9	31.8	37.0	33.2	35.3	33.9	36.5	33.3	37.0	36.7	26.6	23.1
+ fine-tune w/ A	55.7 ↑ 20.8	53.3 ↑ 21.5	53.7 ↑ 16.7	50.4 ↑ 17.2	52.8 ↑ 17.5	52.3 ↑ 18.4	56.5 ↑ 20.0	52.0 ↑ 18.7	56.2 ↑ 19.2	53.1 ↑ 16.4	61.0 ↑ 34.4	60.8 ↑ 37.7
+ fine-tune w/ A&E	54.9 ↑ 20.0	53.0 ↑ 21.2	55.5 ↑ 18.5	51.4 ↑ 18.2	51.9 ↑ 16.6	49.7 ↑ 15.8	55.3 ↑ 18.8	51.5 ↑ 18.2	53.4 ↑ 16.4	51.6 ↑ 14.9	59.3 ↑ 32.7	61.4 ↑ 38.3
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
+ fine-tune w/ A	56.8 ↑ 22.5	55.2 ↑ 21.3	55.2 ↑ 16.1	51.5 ↑ 14.3	59.6 ↑ 26.0	55.2 ↑ 20.9	58.4 ↑ 23.9	54.1 ↑ 18.1	50.5 ↑ 12.0	53.7 ↑ 14.4	61.0 ↑ 38.4	63.7 ↑ 41.0
+ fine-tune w/ A&E	51.0 ↑ 16.7	52.9 ↑ 19.0	48.8 ↑ 9.7	50.8 ↑ 13.6	55.3 ↑ 21.7	51.2 ↑ 16.9	52.2 ↑ 17.7	51.9 ↑ 15.9	44.2 ↑ 5.7	50.1 ↑ 10.8	54.8 ↑ 32.2	60.5 ↑ 37.8

and 2.3% compared to direct answer generation, respectively. We speculate that generating explanations is relatively more challenging, and incorporating it with answer generation might impede the models' capacity to generate correct answers. This finding raises an important question: how can we effectively leverage the interpretability information within PathMMU to enhance models' training?

5 Conclusion

In this study, we introduce PathMMU, the largest and highest-quality pathology benchmark to date, specifically crafted to evaluate the capabilities of LMMs in interpreting and reasoning with pathology images. The construction of PathMMU involves a meticulous data collection and curation process, supplemented by a strict manual review by seven professional pathologists to ensure its quality and professionalism. Moreover, we establish a human expert performance benchmark to quantify the gap between cutting-edge LMMs and human experts. Our experimental results reveal that advanced LMMs significantly lag behind on PathMMU, with these models demonstrating poor performance in observing details in pathology images and sometimes even neglecting visual information, highlighting a substantial gap in practical pathology application. However, LMMs demonstrate notable performance improvements after fine-tuning on PathMMU, even surpassing GPT-4V. While they do not achieve human expert-level performance, these LMMs show promising potential for analyzing pathology images.

Future Directions for Pathology LMMs: Our experience with PathMMU highlights key areas for development in pathology LMMs: (1) Current LMMs, which are primarily based on LLMs and fine-tuned in a lightweight manner, tend to over-rely on textual information while neglecting visual data. There is a significant need to explore training methodologies or model structures that better integrate visual and textual modalities. (2) There's a tendency for LMMs to take shortcuts, solving problems in a "lazy" manner. This necessitates the development of trustworthy models for real-world clinical applications. (3) Given that most current LMMs do not support multi-image inputs, PathMMU does not include a benchmark for processing multiple images. However, in practical scenarios, pathologists typically analyze samples at various magnifications and perspectives, underscoring the importance of developing LMMs capable of handling multi-image inputs. Finally, we believe PathMMU will catalyze significant advancements in the development of next-generation LMMs in pathology.

Acknowledgements

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

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: NeurIPS. pp. 23716–23736 (2022)
2. Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., et al.: Bach: Grand challenge on breast cancer histology images. Medical image analysis **56**, 122–139 (2019)
3. Arunachalam, H.B., Mishra, R., Daescu, O., Cederberg, K., Rakheja, D., Sengupta, A., Leonard, D., Hallac, R., Leavey, P.: 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. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
5. Bavishi, R., Elsen, E., Hawthorne, C., Nye, M., Odena, A., Somani, A., Taşırlar, S.: Introducing our multimodal models (2023), `https://www.adept.ai/blog/fuyu-8b`
6. Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S.A., Müller, H.: Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In: Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes (2021)
7. Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M.: Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142 (2019)
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS. pp. 1877–1901 (2020)
9. Cai, R., Song, Z., Guan, D., Chen, Z., Luo, X., Yi, C., Kot, A.: Benchlmm: Benchmarking cross-style visual capability of large multimodal models. arXiv preprint arXiv:2312.02896 (2023)
10. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), `https://lmsys.org/blog/2023-03-30-vicuna/`
11. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL. pp. 4171–4186 (2019)
13. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. In: ICML. pp. 8469–8488 (2023)

14. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
15. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)
16. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: CVPR. pp. 14375–14385 (2024)
17. Han, C., Pan, X., Yan, L., Lin, H., Li, B., Yao, S., Lv, S., Shi, Z., Mai, J., Lin, J., et al.: Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. arXiv preprint arXiv:2204.06455 (2022)
18. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
19. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
20. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. In: NeurIPS. pp. 37995–38017 (2023)
21. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo10 **5281** (2018)
22. Kriegsmann, K., Lobers, F., Zgorzelski, C., Kriegsmann, J., Janssen, C., Meliss, R.R., Muley, T., Sack, U., Steinbuss, G., Kriegsmann, M.: 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)
23. Kumar, V., Abbas, A.K., Fausto, N., Aster, J.C.: Robbins and Cotran pathologic basis of disease, professional edition e-book. Elsevier health sciences (2014)
24. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
25. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023)
26. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: benseed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
27. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In: NeurIPS. pp. 28541–28564 (2023)
28. Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., et al.: Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 (2022)
29. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML. pp. 19730–19742 (2023)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS. pp. 34892–34916 (2023)

31. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)
32. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR. pp. 11976–11986 (2022)
33. OpenAI: Introducing chatgpt. https://openai.com/blog/chatgpt (2022)
34. OpenAI: Gpt-4 technical report (2023)
35. OpenAI: Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf (2023)
36. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
38. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
39. Seyfioglu, M.S., Ikezogwo, W.O., Ghezloo, F., Krishna, R., Shapiro, L.: Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In: CVPR. pp. 13183–13192 (2024)
40. Silva-Rodríguez, J., Colomer, A., Sales, M.A., Molina, R., Naranjo, V.: 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)
41. Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., et al.: Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137 (2021)
42. Sun, Y., Zhang, Y., Si, Y., Zhu, C., Shui, Z., Zhang, K., Li, J., Lyu, X., Lin, T., Yang, L.: Pathgen-1.6m: 1.6 million pathology image-text pairs generation through multi-agent collaboration (2024), https://arxiv.org/abs/2407.00203
43. Sun, Y., Zhu, C., Zhang, Y., Li, H., Chen, P., Yang, L.: Assessing the robustness of deep learning-assisted pathological image analysis under practical variables of imaging system. In: ICASSP. pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10095887
44. Sun, Y., Zhu, C., Zheng, S., Zhang, K., Sun, L., Shui, Z., Zhang, Y., Li, H., Yang, L.: Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In: AAAI. pp. 5034–5042 (2024)
45. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
46. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
47. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: MICCAI. pp. 210–218. Springer (2018)
48. Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., et al.: Evaluation and analysis of hallucination in large vision-language models. arXiv preprint arXiv:2308.15126 (2023)

49. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023)
50. Wei, J., Suriawinata, A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L., Brown, C., Baker, M., Tomita, N., Torresani, L., 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. pp. 11–24. Springer (2021)
51. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. arXiv preprint arXiv:2306.09265 (2023)
52. Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Huang, X., Wang, Z., Sheng, L., BAI, L., Shao, J., Ouyang, W.: Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In: NeurIPS. pp. 26650–26685 (2023)
53. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
54. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: CVPR. pp. 9556–9567 (2024)
55. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
56. Zhang, Y., Sun, Y., Li, H., Zheng, S., Zhu, C., Yang, L.: Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. In: MICCAI. pp. 242–252 (2022)
57. Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., et al.: Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nature Machine Intelligence **1**(5), 236–245 (2019)
58. Zheng, S., Cui, X., Sun, Y., Li, J., Li, H., Zhang, Y., Chen, P., Jing, X., Ye, Z., Yang, L.: Benchmarking pathclip for pathology image analysis. Journal of Imaging Informatics in Medicine pp. 1–17 (2024)
59. Zhu, C., Sun, Y., Li, H., Cui, C., Zhang, S., Cai, J., Ling, Y.: Weakly supervised classification using multi-level instance-aware optimization on cervical cytologic image. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2022)
60. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)