PathAsst: A Generative Foundation AI Assistant Towards Artificial General Intelligence of Pathology

Yuxuan Sun[1,2,*], Chenglu Zhu[2,*], Sunyi Zheng[2], Kai Zhang[3], Lin Sun[4], Zhongyi Shui[1,2], Yunlong Zhang[1,2], Honglin Li[1,2], Lin Yang[2,†]

[1]College of Computer Science and Technology, Zhejiang University, China
[2]Research Center for Industries of the Future and School of Engineering, Westlake University, China
[3]Department of Computer Science and Engineering, The Ohio State University, USA
[4]School of Computer and Computing Science, Hangzhou City University, China
yanglin@westlake.edu.cn

Abstract

As advances in large language models (LLMs) and multimodal techniques continue to mature, the development of general-purpose multimodal large language models (MLLMs) has surged, offering significant applications in interpreting natural images. However, the field of pathology has largely remained untapped, particularly in gathering high-quality data and designing comprehensive model frameworks. To bridge the gap in pathology MLLMs, we present PathAsst, a multimodal generative foundation AI assistant to revolutionize diagnostic and predictive analytics in pathology. The development of PathAsst involves three pivotal steps: data acquisition, CLIP model adaptation, and the training of PathAsst's multimodal generative capabilities. Firstly, we collect over 207K high-quality pathology image-text pairs from authoritative sources. Leveraging the advanced power of ChatGPT, we generate over 180K instruction-following samples. Furthermore, we devise additional instruction-following data specifically tailored for invoking eight pathology-specific sub-models we prepared, allowing the PathAsst to effectively collaborate with these models, enhancing its diagnostic ability. Secondly, by leveraging the collected data, we construct PathCLIP, a pathology-dedicated CLIP, to enhance PathAsst's capabilities in interpreting pathology images. Finally, we integrate PathCLIP with the Vicuna-13b and utilize pathology-specific instruction-tuning data to enhance the multimodal generation capacity of PathAsst and bolster its synergistic interactions with sub-models. The experimental results of PathAsst show the potential of harnessing AI-powered generative foundation model to improve pathology diagnosis and treatment processes. We open-source our dataset, as well as a comprehensive toolkit for extensive pathology data collection and preprocessing at https://github.com/superjamessyx/Generative-Foundation-AI-Assistant-for-Pathology.

Introduction

In recent years, artificial intelligence has made remarkable strides across various fields (Liu et al. 2022b; Zhuang et al. 2021). This is particularly evident in pathology, which has undergone a profound transformation with the introduction of digital pathology and advanced deep learning techniques. The increasing availability of digitized histopathology data, coupled with the exponential growth in the size and complexity of pathology datasets, has necessitated the development of more sophisticated tools to enhance the analytical efficiency of pathologists.

Simultaneously, there has been an upsurge interest in LLMs, with numerous researchers focusing on their development and application. The ultimate goal is to create models with general artificial intelligence capabilities. Among the most prominent examples are OpenAI's ChatGPT and GPT-4. These models have showcased impressive capabilities in human interaction by training through instruction tuning and human feedback, thereby fueling the community's enthusiasm for LLMs.

In the open-source community, LLaMA (Touvron et al. 2023) has emerged as a compelling model that exhibits performance on par with GPT-3 (Brown et al. 2020), providing promising opportunities for further development. Subsequent models, such as Alpaca (Taori et al. 2023) and Vicuna (Chiang et al. 2023), take advantage of LLaMA and leverage the instruction tuning techniques, enabling them even outperform ChatGPT in certain tasks. Researchers have also explored the realm of multimodal models, creating innovative approaches such as LLaVA (Liu et al. 2023a) and MiniGPT-4 (Zhu et al. 2023). These models demonstrate impressive capabilities in comprehending and interpreting multimodal data, showcasing the advancements in the field.

However, while these advanced MLLMs primarily focus on natural images, the field of pathology faces a notable gap due to the scarcity of high-quality data and limited exploration of model frameworks, which results in a deficiency of pathology-specific MLLMs. In this study, we aim to bridge this gap by exploring both high-quality pathology data collection and the potential application of MLLMs within the pathology domain. We outline our contributions as follows:

- We gather diverse pathology image-caption pairs from authoritative sources. Through a meticulous process of data cleaning and optimization, we create the PathCap dataset, comprising 207K high-quality samples.
- We introduce PathCLIP, a pathology-specific CLIP

*These authors contributed equally.
†Corresponding Author.

model trained on the PathCap. Compared to prior models, PathCLIP shows superior proficiency in understanding pathology data, achieving state-of-the-art results in pathology image retrieval and zero-shot classification.

- We integrate PathCLIP and Vicuna-13b to develop PathAsst, a multimodal generative foundational model tailored for pathology. Utilizing the PathCap dataset, we prompt ChatGPT to generate the PathInstruct dataset, which consists of 180K pathology multimodal instruction-following samples. These samples are employed to train PathAsst's generative capabilities. Additionally, we prepare eight pathology-specific sub-models, supplemented with instruction-following data for various scenarios that necessitate sub-model invocation. This equips PathAsst with the ability to discern when to utilize these models for optimal results.

Related Work

Large Language Model (LLM). In the early stages, breakthrough models like BERT (Devlin et al. 2018) and GPT (Radford et al. 2018), were introduced, drawing inspiration from the transformer architecture. These models ignited significant interest in the natural language processing (NLP) domain and signaled the beginning of large-scale models in this field. Initially, the full potential of generative models remained largely unexplored. However, in recent years, as the generative model continue to scale up, more powerful models such as GPT-3 (Brown et al. 2020), T5 (Raffel et al. 2020), PaLM (Chowdhery et al. 2022), and OPT (Zhang et al. 2022) are developed. Their emergent abilities (Wei et al. 2022) lead these larger models to display markedly superior performance on complex tasks compared to their smaller counterparts. Furthermore, the introduction of instruction tuning techniques (Ouyang et al. 2022; Wang et al. 2022b,a), specifically in the realm of LLM, enables the generation of more controllable, practical, and task-specific results. This revolutionary enhancement significantly boosts the zero-shot learning abilities of large models, as exemplified by InstructGPT (Ouyang et al. 2022), GPT-4 (OpenAI 2023), FLAN-T5 (Chung et al. 2022), and FLAN-PaLM (Chung et al. 2022).

Multimodal Large Language Model (MLLM). Recent advancements in large-scale multimodal models can be primarily divided into two branches. The first branch is developed based on the LangChain (Chase 2022) approach, where LLM collaborates with various specialized visual models to generate results. Prominent representatives of this branch include Visual ChatGPT (Wu et al. 2023) and MM-REACT (Yang et al. 2023). The second branch is implemented by integrating the feature outputs from visual models into the token sequence inputs of the LLM, enabling multimodal generation. This method is represented in models such as BLIP-2 (Li et al. 2023), PaLM-E (Driess et al. 2023) and Flamingo (Alayrac et al. 2022). Building upon the instruction-tuning techniques inspired by the LLM community, researchers create multimodal instruction-following datasets to perform MLLM training. This approach promptes the development of models such as LLaVA (Liu et al. 2023a), MiniGPT-4 (Zhu et al. 2023) and LLaMA-Adapter V2 (Gao et al. 2023). These models demonstrate impressive performance in solving multimodal tasks, as well as advanced multimodal chat capabilities.

Multimodal Model for Pathology. While there are numerous applications for multimodal models in natural image analysis, their use in pathological image analysis has been relatively limited to date. The majority of methods employ approaches that combine vision encoder with LSTM (Liu et al. 2023b; Zhang et al. 2019a,b), yielding fairly satisfactory results. TraP-VQA (Naseem, Khushi, and Kim 2022) is the first attempt to employ vision-language transformer in pathology image processing, which is tested on the PathVQA dataset (He et al. 2020) to generate interpretable answers. More recently, Huang et al. (Huang et al. 2023) compile a large-scale dataset of pathology image-text pairs, sourced from social media platforms such as Twitter. They utilize contrastive vision-language pretraining to establish a foundational model for pathology, demonstrating promising results in pathology zero-shot image-text cross-modal retrieval and zero-shot image classification.

Multimodal Datasets. Numerous researchers have been dedicating their efforts to contribute valuable datasets that facilitate the advancement of models in the aforementioned domains. For instance, in the general domain, the community has successfully constructed various datasets, such as CC (Changpinyo et al. 2021) and LAION (Schuhmann et al. 2022). In the biomedical field, researchers have released datasets like ROCO (Pelka et al. 2018), MedICAT (Subramanian et al. 2020), and PMC-OA (Lin et al. 2023). In the pathology domain, researchers have recently built the OpenPath (Huang et al. 2023) dataset by crawling Twitter.

Despite significant progress in the field, the domain of MLLM specifically adapted for pathology remains largely untapped. Current models, primarily designed for caption generation, often underperform when compared to specialized professional pathology models. Furthermore, regarding pathology MLLM dataset construction, existing datasets such as ROCO, MedICAT, and PMC-OA are not specifically tailored for this field. The only large-scale dataset, OpenPath, primarily sources its data from Twitter, where the image-text correlation is relatively weak, thus posing challenges for MLLM training. Moreover, the image-text pairs in OpenPath require access to the Twitter API, which carries a significant cost. As a result, there is still a substantial lack of high-quality image-caption datasets in the field of pathology. To bridge this gap, we develop two comprehensive pathology multimodal datasets. Building on these datasets, we utilize the power of instruction tuning to significantly improve MLLM's capability in interpreting pathology images.

Pathology Dataset Construction

In this paper, we propose two datasets tailored for pathology: **PathCap** and **PathInstruct**. The PathCap contains 207K high-quality pathology image-caption pairs. Among them, 197K are collected from PubMed and internal pathology guidelines books, while an additional 10K annotations are

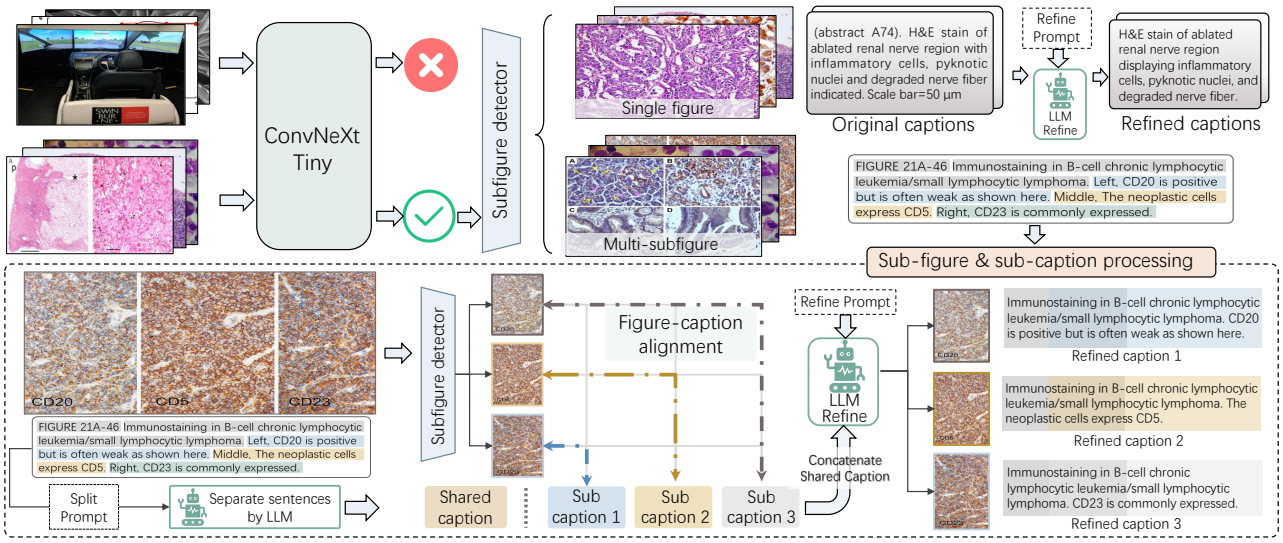

Figure 1: Illustration of data processing: pathology image selection, sub-figure & caption separation, and refinement.

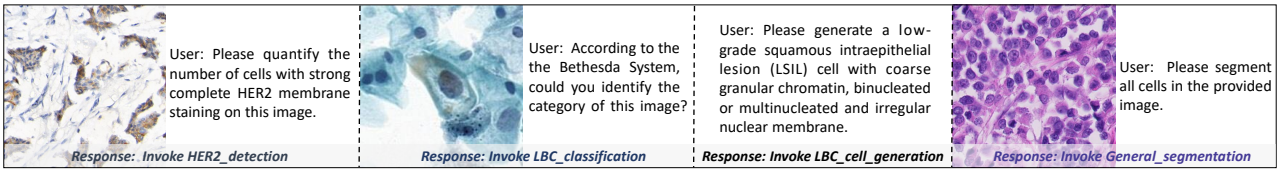

Figure 2: Examples of pathology-specific model-invoking instruction-following samples.

provided by expert cytologists specializing in liquid-based cytology (LBC). The PathInstruct dataset consists of 180K samples and includes two parts of instruction-following data. The first part is generated by prompting ChatGPT based on curated pathology image-text pairs (refer to step 4 in the subsequent data processing introduction). The second section includes multimodal instruction-following data tailored for model invocation, ensuring the effective use of specialized pathology models based on user intent and image features.

More specifically, data from PubMed are parsed from XML format papers into image-text pairs. For books, we first convert them from PDF to HTML and then parse the content into image-text pairs. Through these efforts, we collect 15M and 2K samples from these respective sources. Although the amount of data available on PubMed is substantial, it should be noted that the proportion of the data related to pathology is limited. Additionally, the clarity of these pathology images is comparatively inferior. Therefore, thorough filtering is required to ensure the quality and relevance of image-text pairs. As shown in Figure 1, our data cleansing process is executed methodically, following four carefully designed steps:

Step 1: Pathology data selection. The dataset collected, especially from PubMed, encompasses a wide variety of image sources beyond the scope of pathology. To efficiently select pathology-related data, we manually annotate 20K samples, categorizing them as either pathological or non-pathological. Subsequently, we train a ConvNeXt (Liu et al. 2022a) model to identify pathological data within the remaining dataset, resulting in a pathology-specific dataset comprising 135K pathology-specific images.

Step 2: Sub-figure and sub-caption separation & alignment. In many instances, images consist of multiple sub-figures, necessitating precise separation and alignment with their corresponding captions. As depicted in the lower half of Figure 1, we address the sub-figure separation by developing a YOLOv7 model (Wang, Bochkovskiy, and Liao 2022) trained on 2K annotated bounding boxes. Regarding caption separation, conventional rule-based methods often fail to handle the separation of diverse and intricate captions. To overcome this limitation, we leverage the power of ChatGPT to automatically separate approximately 60K captions using carefully crafted prompts. Subsequently, we employ PLIP (Huang et al. 2023) to align sub-image with its corresponding sub-caption by assessing the similarity of visual content and captions. Moreover, we eliminate images with lower resolution, and remove the less relevant image-text pairs, further enhancing the overall quality of the dataset. Ultimately, we acquire 195K high-quality image-text pairs.

Step 3: Caption refinement. As original captions include irrelevant information such as age and disease descriptions, and are not presented in a descriptive style. We design prompts to employ ChatGPT in refining the captions, making them more suitable for training.

Step 4: Instruction-following data generation. In this

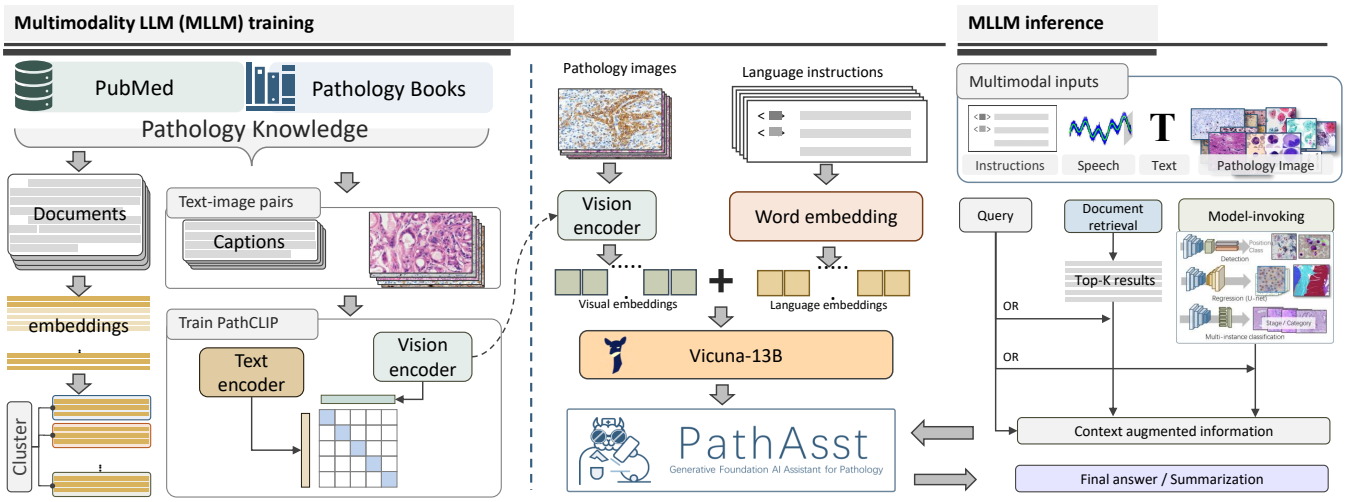

Figure 3: An illustration of the overall framework of PathAsst. The multimodal MLLM training encompasses the training processes of both PathCLIP and PathAsst, as well as the construction of a paper embedding database. The tool-augmented MLLM inference details the process of PathAsst utilizing various tools to enhance the quality of its generated outputs.

step, we select image-text pairs with captions exceeding 12 words. Using these pairs, we produce two types of instruction-following data: detailed description-based and conversation-based. The former is created by applying multiple well-designed instructions that inquire about detailed information, while the latter involves using ChatGPT to generate conversational Q&As based on the captions. Additionally, we design special model-invoking instruction-following samples covering a diverse range of scenarios, as depicted in Figure 2, enabling PathAsst with the capability to appropriately utilize pathology-specific sub-models.

PathAsst Framework Construction

In this section, we present a comprehensive description of the construction process of PathAsst. This includes the introduction to the design of the model's structure, training methodology, and the tools used for augmented model inference. A general overview can be found in Figure 3.

Model Design and Training

PathAsst is designed to integrate the strengths of both the advanced LLM and the CLIP (Radford et al. 2021) vision encoder to enable enhanced pathological analysis. For the visual component, we employ our custom-trained PathCLIP, complemented by a fully connected (FC) layer. Concerning the LLM component, we utilize Vicuna-13B (Chiang et al. 2023), a model widely recognized as the closest to ChatGPT in terms of performance. To elaborate, when an input image is provided, it is first encoded into visual tokens via the PathCLIP. Subsequently, the FC layer maps the image embedding space to the corresponding language embedding space. Finally, both visual and language embeddings are concatenated to the inputs of the MLLM. In the following, we introduce the detailed training process of PathCLIP and PathAsst.

Training of PathCLIP. As one of the core components of PathAsst, the capability of CLIP in interpreting pathological images largely dictates the performance ceiling of PathAsst. Therefore, we develop PathCLIP, a specialized variant of CLIP tailored for pathology. The training process involves fine-tuning a pre-trained OpenAI CLIP base model (Radford et al. 2021) using our PathCap dataset in a contrastive learning approach, following the training procedure from OpenCLIP repository (Ilharco et al. 2021). To be specific, for a batch of N image-text pairs, PathCLIP is designed to maximize the cosine similarity between the embeddings of the pathology image and its corresponding text within each batch. Concurrently, it minimizes the cosine similarity amongst the remaining $N^2 - N$ non-pair samples. This strategy aligns the pathology vision and language space, thereby endowing PathCLIP with a more effective interpretation and analysis of pathology images.

Training of PathAsst. PathAsst is trained using the PathInstruct dataset through a two-phase training. In the first phase, both the vision encoder and the LLM are frozen, and we only train the FC layer that connects to the vision encoder. This initial phase aims to preliminarily align the vision encoder with the LLM. During this phase, we utilize the detailed description-based part of the PathInstruct. In the second phase, with an aspiration for PathAsst to generate higher-quality and more detailed responses, we extract all the data from books within the PathInstruct dataset, and include samples from PubMed with single images and captions exceeding the length of 50 tokens, resulting in a total training set of 35K samples. Only the PathCLIP is frozen during this phase's training.

Specifically, we standardize both forms of instruct-following data formats, as shown in Table 1. First, we predefine a system message that sets the context for the LLM role. This is followed by a conversation between the user and the assistant, where the user provides instructions, and the assis-

$X_{\texttt{system-message}}$ <STOP> \n
User: <image_token>\n {instruction} <STOP> \n
Assistant: {response} <STOP> \n
User: {instruction} <STOP>\n
Assistant: {response} <STOP> \n .....

Table 1: Illustration of instruction-following data format, where {instruction} represents the user query, {response} denotes the corresponding answer. The $X_{\texttt{system-message}}$ is set as: A dialogue between a professional pathology assistant and a human. The assistant provide informative, helpful, and detailed answers. The <STOP> is represented by ###, while <image_token> stands for the tokens corresponding to the image tokens. During the model training, only {response} is considered when calculating loss.

tant responds accordingly based on the instructions. To fine-tune our model, we utilize instruction-tuning via next-word prediction. Specifically, the model is trained to optimize the likelihood of generating an accurate response given the input image $\mathcal{I}$ and instruction $\mathbf{X}_{instruct}$. The loss is calculated using the negative log-likelihood of the correct next token in the sequence, with the total loss summed across all time steps, which can be formulated as:

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log p\left(x_t \mid \mathcal{I}, \mathbf{X}_{instruct}, \mathbf{X}_{\mathrm{a},<t}; \boldsymbol{\theta}\right), \quad (1)$$

Where $\mathbf{X}_{\mathrm{a},<t}$ refers to the prior tokens in the response sequence, θ denotes the trainable parameters of PathAsst. Specifically, during the first phase of training, θ corresponds to the parameters of the FC layer. In the subsequent phase, it represents both FC layer and LLM parameters. Meanwhile, T signifies the length of the ground-truth response, and $p\left(x_t \mid \mathcal{I}, \mathbf{X}_{instruct}, \mathbf{X}_{\mathrm{a},<t}; \boldsymbol{\theta}\right)$ represents the probability of generating the t-th token in the response sequence.

Tool Augmented MLLM Inference

To augment PathAsst's capabilities and offer more precise responses, we prepare two types of tools that PathAsst can employ during its inference phase. One leverages pathology-specific computer vision (CV) sub-models, while the other focuses on paper retrieval. These tools not only enrich the context for PathAsst but also enable tasks beyond text generation, such as image generation and segmentation.

Pathology-specific CV Model Zoo. We integrate eight specialized pathological models into PathAsst for seamless invocation: **(1)** LBC (liquid-based cytology) classification model: This model is based on ConvNeXt-Tiny (Liu et al. 2022a), specifically designed for liquid-based cervical cytology image classification. Through the analysis of abnormal cell morphologies within the image, it effectively classifies the image into one of the six categories as defined by The Bethesda System (TBS). **(2)** LBC detection model: We utilize YOLOv7 (Wang, Bochkovskiy, and Liao 2022) as the backbone for developing our detection model, which is employed to identify abnormal cells within image patches. This model is specifically designed to detect the five classes of non-normal cells as defined in TBS. **(3)** Hematological cell detection model: This model, developed based on YOLOv7, specializes in blood cell classification, which is crucial for diagnosing various hematological conditions. **(4)** LBC cell generation model: This model is developed based on Stable Diffusion (Rombach et al. 2022), which is capable of generating specific cells based on user input, such as 'generate an image of a cell with nuclei enlarged 2-2.5 times'. **(5)** HER2 detection model, **(6)** PD-L1 detection model and **(7)** Ki67 detection model are developed using DPA-P2PNet (Shui et al. 2023) for immunohistochemical cell detection and classification. **(8)** General segmentation model: Benefiting from the outstanding general segmentation quality of the Segment Anything Model (Kirillov et al. 2023), we directly employ it as our pathology image segmentation model.

Once PathAsst invokes a particular specialized model, it processes both the user's query and the output of the invoked model to formulate a conclusive response, resulting in a more precise and effective interaction with the user.

Model	CRC	WSSS4LUAD	LC-lung	LC-colon
OpenAI CLIP	22.2	61.6	31.5	75.7
PLIP	53.1	69.5	86.0	87.0
PathCLIP	**54.2**	**81.1**	**88.7**	**94.3**

Table 2: Comparative evaluation of zero-shot image classification performance across different CLIP models.

Enhancing Responses through Paper Retrieval. In the realm of pathology, even the highly recognized GPT-4 struggles with specific queries that necessitate deep domain knowledge, especially apparent when addressing questions involving the most recent research. Taking inspiration from Langchain's approach (Chase 2022) for building local knowledge databases, we gather 5.3M article abstracts from PubMed. We utilize PubMedBERT (Gu et al. 2021) for abstract embedding extraction and Faiss (Johnson, Douze, and Jégou 2019) for the efficient storage of these embeddings. To expedite inference efficiency, a preliminary abstract clustering is conducted. Upon user query, our system allows the extraction of relevant information from this paper database, serving as context information to amplify the precision of LLM's responses.

Experiments

Evaluation Datasets Construction. We construct and gather a series of test datasets to evaluate the performance of the proposed PathCLIP and PathAsst.

For the evaluation of zero-shot classification of PathCLIP, we collect: (1) CRC100K dataset (Kather, Halama, and Marx 2018): This is a collection of 100K image patches derived from H&E stained histological images of both colorectal cancer and normal tissue, categorized into nine tissue classes, including Adipose, Background, Debris, Lymphocytes, Mucus, Smooth Muscle, Normal Colon Mucosa, Cancer-Associated Stroma, and Colorectal Adenocarcinoma Epithelium. (2) WSSS4LUAD (Han et al. 2022): This

Dataset	Number of candidates	Metric	OpenAI CLIP	PLIP	PathCLIP	Random	Fold change (PathCLIP VS.PLIP)	Fold change (PathCLIP VS.OpenAI CLIP)
PubMed	12430	R@10	3.1	3	33.2	0.1	11.07	10.71
		R@50	8	8.9	56.9	0.3	6.39	7.11
Books	973	R@10	7.2	17.5	41.6	0.9	2.38	5.78
		R@50	22.7	45.8	72.7	5.9	1.59	3.20

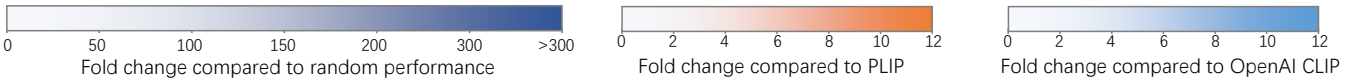

Figure 4: Comparative assessment of image retrieval performance between CLIP models across collected datasets.

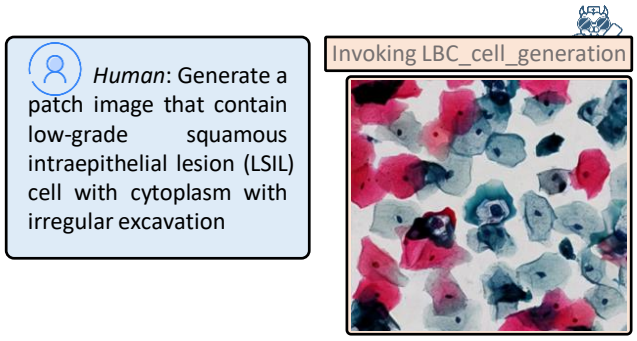

Figure 5: Example of PathAsst calls generation model.

Method	PathVQA	
	Closed	Open
M2I2 (Li et al. 2022)	88.0	36.3
CLIP-ViT w/ GPT2 (van Sonsbeek et al. 2023)	87.0	**40.0**
MMQ (Do et al. 2021)	84.0	13.4
LLaVA (Liu et al. 2023a)	81.0	19.2
BLIP-2 Flan-T5 XXL (Li et al. 2023)	80.1	34.1
PathAsst (w/ CLIP)	89.7	37.6
PathAsst (w/ PathCLIP)	**90.9**	38.4

Table 3: Comparison of various methods on PathVQA.

dataset comprises patch-level annotations from 87 whole slide images. In this case, we focus on the tumor and normal classes, which yields a total of 6,579 tumor and 1,832 normal images. In order to assess the model on these datasets, labels are transformed into complete sentences. For instance, the label 'tumor' is rephrased as 'A H&E image of a tumor.' (3) LC25000 (Borkowski et al. 2019). This dataset comprises tissue samples from lung and colon adenocarcinomas, divided into two distinct subsets: the LC-lung and the LC-colon. The LC-lung encompasses 15,000 images and includes classifications of lung adenocarcinomas, lung squamous cell carcinomas, and benign lung tissues. On the other hand, the LC-colon subset, containing 10,000 images, is categorized into colon adenocarcinomas and benign colonic tissues. The F1 score is used as the metric for evaluation.

For the cross-modal retrieval validation of PathCLIP, we employ the test set from the PubMed section of our collected data, along with data from books. Note that the data from books are not included during the training phase of the PathCLIP, hence providing an evaluation in the context of unseen domains. Given that the length of some captions in these datasets is relatively long and may exceed the token length limitation of CLIP, we opt for samples with captions that are fewer than 77 tokens. The $R@k$ metric is used to assess the performance of image retrieval, which measures whether the correct image is presented among the $Topk$ retrieved images.

For the validation of PathAsst, we employ the PathVQA dataset (He et al. 2020), which comprises 32,799 questions derived from 4,998 pathology images. The type of questions includes open-ended questions typically beginning with what, where, and when, as well as close-ended questions requiring yes/no responses. We measure model performance of close-ended questions using accuracy, and evaluate open-ended questions with F1-score.

Statistical Results. As shown in Table 2 and Figure 4. Our analysis demonstrates that our PathCLIP significantly surpasses the baseline OpenAI CLIP model, consistently outperforming the state-of-the-art (SOTA) pathology model, PLIP, in tasks such as cross-modal image retrieval and zero-shot image classification. To be specific, PathCLIP achieves a remarkable improvement in the $R@10$ retrieval on the PubMed dataset, with a 10.71-fold and 11.07-fold increase compared to the OpenAI CLIP and PLIP models, respectively. In the context of unseen domain data, the retrieval $R@10$ on the books is 5.78 times and 2.38 times that of OpenAI CLIP and PLIP, respectively. Considering the zero-shot classification tasks, PathCLIP achieves a substantial improvement in F1-score compared to CLIP, with notable gains of 32%, 19.5%, 57.2%, and 18.6% on the CRC100K, WSSS4LUAD, LC-lung, and LC-colon datasets, respectively. Furthermore, even when compared to the previous SOTA PLIP model, PathCLIP shows an increase of 1.1%, 11.6%, 2.7%, and 7.3% on these datasets, respectively. For the evaluation on PathVQA, PathAsst significantly outperforms the prior MLLM model in both closed-form and open-ended question types. Specifically, it surpasses LLaVA by 8.7% and 18.4% in these two question types, respectively. This underscores the importance of training with PathInstruct data. Further enhancements of 1.2% and 0.8% are noted after substituting CLIP with PathCLIP, indicating that

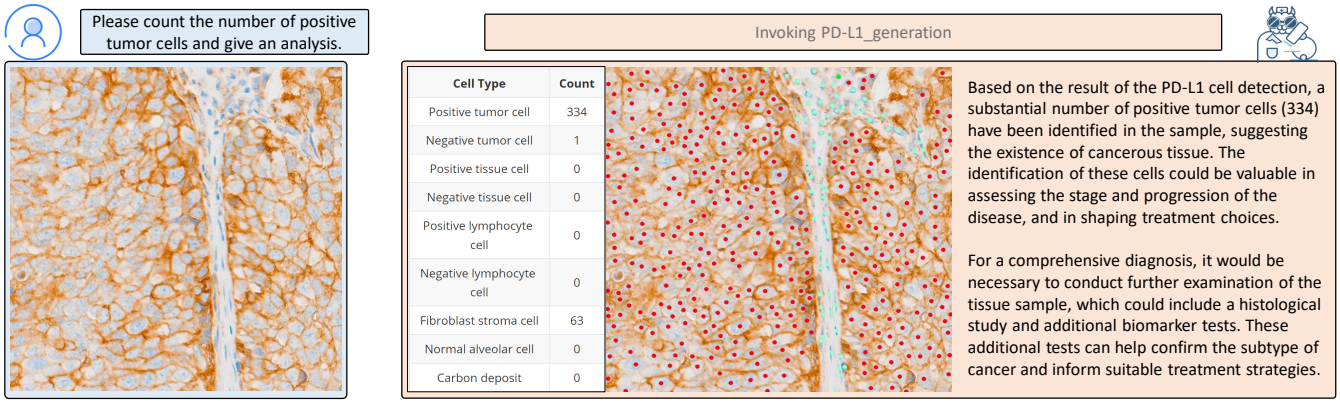

Figure 6: An example of PathAsst invokes the PD-L1 detection model for assistance.

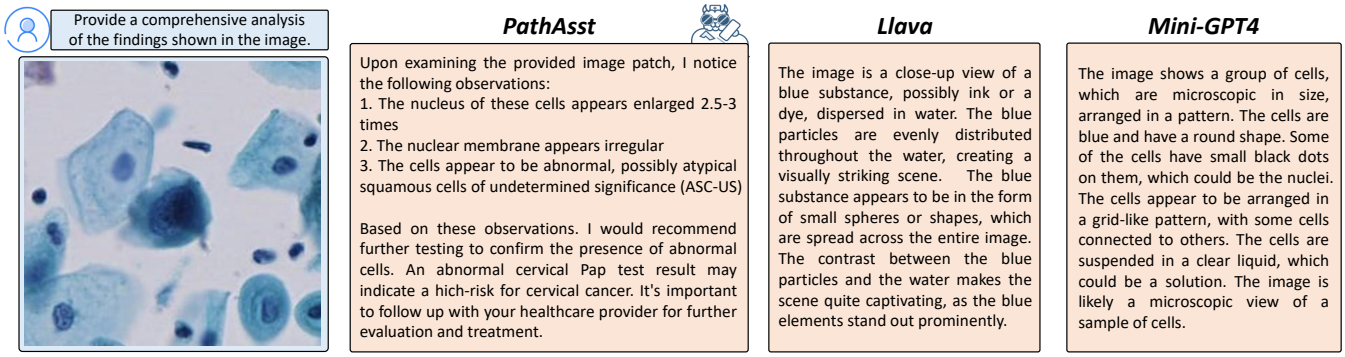

Figure 7: An example of PathAsst, LLaVA, and MiniGPT-4's capability in interpreting pathology images.

the incorporation of PathCLIP enhances PathAsst's understanding of pathology images. Compared with the performance of the previous SOTA model, which directly extracts the statistical number from their reports, PathAsst achieves considerable improvements in closed-ended questions, although it slightly underperforms the SOTA model in open-ended questions.

Demonstration Showcase of PathAsst. Here, we showcase several examples of PathAsst's robust capabilities in handling complex pathology tasks. As shown in Figure 5, PathAsst is capable of recognizing the user's need to generate an LBC cell that belongs to the LSIL category with irregular excavated cytoplasm. It accomplishes this by invoking the LBC cell generation model. This advanced functionality empowers users to create a diverse range of LBC cells that are precisely tailored to their specific needs.

Figure 6 illustrates another example of PathAsst employing a model invocation, where the user requires to count the positive cells in the image, which can be challenging through direct multimodal generation. Therefore, PathAsst chooses to invoke the PD-L1 cell detection model. It automatically marks the predicted points on the cells in the image and provides the statistical results for further analysis with LLM. In this case, LLM generates a markdown-formatted table to display the results along with the corresponding analysis.

Furthermore, Figure 7 demonstrates PathAsst's ability to interpret pathology images independently. In comparison to LLaVA and MiniGPT-4, PathAsst places greater emphasis on cell morphology and features, such as enlarged nucleus and irregular nuclear membrane. In contrast, LLaVA fails to recognize the image as pathological, while MiniGPT-4 generates simplistic descriptions such as 'cells are blue and have a round shape' and 'cells are suspended in a clear liquid.'

Conclusion

In this study, we construct PathCap and PathInstruct datasets, comprising 207K pathology image-text pairs and 180K instruction-following samples, by systematically collecting and processing pathology data from various sources. Leveraging these high-quality datasets, we propose PathCLIP and PathAsst. PathCLIP exhibits powerful capabilities in pathology cross-modal retrieval and zero-shot classification. PathAsst, an instruction-tuned foundation model, is a synergy of the powerful vision encoder PathCLIP and the Vicuna-13b LLM, equipped with an established toolkit that includes eight pathology-specific models and a 5.3 million-sized paper retrieval system. PathAsst not only showcases impressive capabilities in pathology multimodal dialogue and interpreting pathology images, but also the ability to handle more complex pathology tasks by the invocation of these established pathology tools. We hope that the construction of model frameworks and datasets can offer insights and aid in the advancement of pathology foundational models.

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), and the Research Center for Industries of the Future (RCIF) at Westlake University.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*.

Borkowski, A. A.; Bui, M. M.; Thomas; et al. 2019. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3558–3568.

Chase, H. 2022. LangChain.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Do, T.; Nguyen, B. X.; Tjiputra, E.; Tran, M.; Tran, Q. D.; and Nguyen, A. 2021. Multiple meta-model quantifying for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, 64–74. Springer.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010*.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23.

Han, C.; Pan, X.; Yan, L.; Lin, H.; Li, B.; Yao, S.; Lv, S.; Shi, Z.; Mai, J.; Lin, J.; et al. 2022. WSSS4LUAD: Grand Challenge on Weakly-supervised Tissue Semantic Segmentation for Lung Adenocarcinoma. *arXiv preprint arXiv:2204.06455*.

He, X.; Zhang, Y.; Mou, L.; Xing, E.; and Xie, P. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Huang, Z.; Bianchi, F.; Yuksekgonul, M.; Montine, T.; and Zou, J. 2023. Leveraging medical Twitter to build a visual–language foundation model for pathology AI. *bioRxiv*, 2023–03.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP.

Johnson, J.; Douze, M.; and Jégou, H. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3): 535–547.

Kather, J. N.; Halama, N.; and Marx, A. 2018. 100,000 histological images of human colorectal cancer and healthy tissue.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, P.; Liu, G.; Tan, L.; Liao, J.; and Zhong, S. 2022. Self-supervised vision-language pretraining for Medical visual question answering. *arXiv preprint arXiv:2211.13594*.

Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. PMC-CLIP: Contrastive Language-Image Pre-training using Biomedical Documents. *arXiv preprint arXiv:2303.07240*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Liu, H.; Peng, L.; Xie, Y.; Li, X.; Bi, D.; Zou, Y.; Lin, Y.; Zhang, P.; and Li, G. 2023b. Describe like a pathologist: Glomerular immunofluorescence image caption based on hierarchical feature fusion attention network. *Expert Systems with Applications*, 213: 119168.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022a. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Liu, Z.; Wu, S.; Jin, S.; Ji, S.; Liu, Q.; Lu, S.; and Cheng, L. 2022b. Investigating pose representations and motion contexts modeling for 3D motion prediction. *IEEE transactions*

on pattern analysis and machine intelligence, 45(1): 681–697.

Naseem, U.; Khushi, M.; and Kim, J. 2022. Vision-language transformer for interpretable pathology visual question answering. *IEEE Journal of Biomedical and Health Informatics*.

OpenAI. 2023. GPT-4 Technical Report.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Pelka, O.; Koitka, S.; Rückert, J.; Nensa, F.; and Friedrich, C. M. 2018. Radiology Objects in COntext (ROCO): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, 180–189. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

Shui, Z.; Zheng, S.; Yu, X.; Zhang, S.; Li, H.; Li, J.; and Yang, L. 2023. Deformable Proposal-Aware P2PNet: A Universal Network for Cell Recognition under Point Supervision. *arXiv preprint arXiv:2303.02602*.

Subramanian, S.; Wang, L. L.; Mehta, S.; Bogin, B.; van Zuylen, M.; Parasa, S.; Singh, S.; Gardner, M.; and Hajishirzi, H. 2020. Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

van Sonsbeek, T.; Derakhshani, M. M.; Najdenkoska, I.; Snoek, C. G.; and Worring, M. 2023. Open-ended medical visual question answering through prefix tuning of language models. *arXiv preprint arXiv:2303.05977*.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2022a. Self-Instruct: Aligning Language Model with Self Generated Instructions. *arXiv preprint arXiv:2212.10560*.

Wang, Y.; Mishra, S.; Alipoormolabashi, P.; Kordi, Y.; Mirzaei, A.; Arunkumar, A.; Ashok, A.; Dhanasekaran, A. S.; Naik, A.; Stap, D.; et al. 2022b. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; and Duan, N. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Yang, Z.; Li, L.; Wang, J.; Lin, K.; Azarnasab, E.; Ahmed, F.; Liu, Z.; Liu, C.; Zeng, M.; and Wang, L. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, Z.; Chen, P.; McGough, M.; Xing, F.; Wang, C.; Bui, M.; Xie, Y.; Sapkota, M.; Cui, L.; Dhillon, J.; et al. 2019a. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1(5): 236–245.

Zhang, Z.; Chen, P.; Shi, X.; and Yang, L. 2019b. Text-guided neural network training for image recognition in natural scenes and medicine. *IEEE transactions on pattern analysis and machine intelligence*, 43(5): 1733–1745.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhuang, Y.; Liu, Z.; Qian, P.; Liu, Q.; Wang, X.; and He, Q. 2021. Smart contract vulnerability detection using graph neural networks. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 3283–3290.