LLaVA-PlantDiag: Integrating Large-scale Vision-Language Abilities for Conversational Plant Pathology Diagnosis

Karun Sharma
Bennett University
mailto.karun.py@gmail.com

Vidushee Vats
Bennett University
vatsvidushee@gmail.com

Abhinendra Singh
Bennett University
e21cseu0039@bennett.edu.in

Rahul Sahani
Bennett University
rahul.sahan810@gmail.com

Deepak Rai
Bennett University
deepak.rai@bennett.edu.in

Ashok Sharma
University of Jammu
Ashoksharma@jammuuniversity.ac.in

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10651096

***Abstract*—Human X Machine Conversational Systems and Multi-turn Generative AI models have established impressive results making the process of language generation and conversation highly interactive in nature. Owing to its versatile functionalities, the models have catered to diverse fields, one of them being Botany and Plant Life. These advancements have contributed and aided phytologists to carry out further research by identifying the plant disease in order to extract other vital information. The current research methods have accomplished the same by implementing the traditional Convolution Neural Networks(CNN's) which not only fail to capture temporal information but limits the interactivity of the user with the model. In this paper, we propose LLaVA-PlantDiag, a visual question-answering assistant which analyzes the data of multiple modalities and answers open-ended questions on plant pathology in a conversational manner. The primary notion is to construct a VQA image-description dataset from the PlantVillage dataset, utilize GPT-3.5 to form open-ended, caption-connected question-answer pair and fine-tune large general-domain vision-language model on the custom dataset. The results demonstrate that LLaVA-PlantDiag significantly outperforms state-of-the-art models such as GPT-4 Vision, Gemini, and other open-source models in two key tasks: Phytopathological multi-turn VQA and Classification. LLaVA-PlantDiag achieves a relative score of 64.7, surpassing the score of 48.7 achieved by GPT-4 Vision on Vision-Language tasks. LLaVA-PlantDiag also obtains an impressive 96% accuracy in classification, compared to the second-best performer, IDEFICUS, which scored 85%.**

***Index Terms*—Multimodal, LLM, LLaVA, Phytopathological Multimodal Data**

I. Introduction

Researchers and scientists are transforming Artificial Intelligence(AI) with the goal of bringing AI to everyone. From research that extends the horizons of this possibility, it is important to involve the integration of human interaction with their surroundings into intelligent machines to set a new threshold of progress that satisfies the human goal of accomplishing real-world assignments in dynamic settings [1].

With the emergence of various multimodal models in the field of visual question answering, text-image generation and natural language for visual reasoning, large language multi-modality models like the LLaVa [1] and GPT-4 have exhibited human-like competence by mimicking human-aptitude and intelligence to the closest. Fine-tuning these models to accord with the multimodalities produce strong performance on the traditional class-level zero-shot tasks as well as the new task-level zero-shots which generalise on the study of unknown datasets [2], [3].

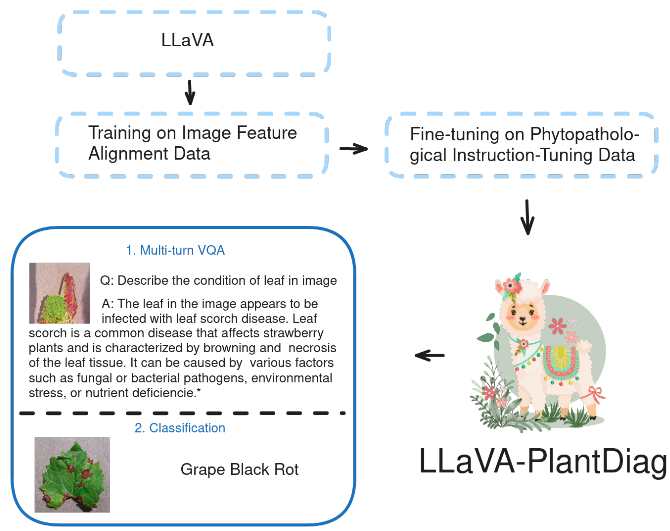

Fig. 1: LLaVA-PlantDiag is initialized from base LLaVA and then adapted to Phytopathological data using 2-stage learning process.

While working on a generalised task, a large language model (LLM) may be a feasible option, but its effectiveness diminishes when the scope is narrowed down to focus on a specific domain [4], [5]. In our context, the focus lies within the field of phytopathology. In our experimentation, the application of the general-task-based LLaVA, OpenAI's GPT-

4 Vision, and Gemini were extended to the phytopathological domain. The results revealed that the model struggled to produce precise responses, offering generalised solutions that applied universally to all instances of leaves. For example, as observed in TABLE I, for an input image of strawberry leaf scorch as shown in Fig 2, the current models generate hallucinations and generalized responses.

Therefore, the current research proposes to synthesise the development of LLaVA-PlantDiag, a Large Language and Vision Assistant for plant diagnosis, using the large language vision model LLaVA as its base model. LLaVA-PlantDiag is structured by a two-stage process that involves the alignment of features via pre-training and fine-tuning, with the architecture outlined in Fig 1. Pre-training the model allows enhanced high-quality graphical representation. In particular, a pre-trained large language model results not only in potent zero-shot capabilities but also strong language generation [6]. LLaVA-PlantDiag is fine-tuned on a custom multimodal phytopathological dataset, stemming from the PlantVillage dataset.

According to the citations from Google Scholar, PlantVillage dataset is the most frequently used dataset for carrying out research in plant study [7]. Most of the research carried out until now uses the PlantVillage dataset for minimal unimodal tasks like segmentation, classification and detection. We harness the data from PlantVillage dataset to generate question-answer pairs obtained from the extended caption formed by GPT-3.5 to form a conversation.

Motivated by recent advancements in [1]–[5] and recognizing their limitations in offering a non-generalized answers for the plant phytopathology leaf image input, We introduce LLaVA-PlantDiag, a domain-specific conversational assistant, specialised in Plant Pathology, fine-tuned on a custom dataset with PlantVillage dataset as base, by using an original instructional approach that has the capability to act as a visual question-answering assistant, equipped to answer open-ended questions and engage in a conversation on plant science and phytology [8]–[10].

The key contributions of the paper are as follows:

1) Phytopathological multimodal instruction-following data: We present an inventive and original data pipeline from the PlantVillage dataset by randomly sampling 200 images from each class. The textual captions are passed into GPT-3.5 to generate extended descriptions, which are then used to generate question-answer conversation pairs. The following is done by exclusively processing text without the usage of image at any stage.
2) LLaVA PlantDiag: We demonstrate an original instructional approach for modulating LLaVA to phytopathological domain, adapted specifically to deal with plant disease and life by using a custom-fit phytopathological multimodal instruction-following dataset. During Feature alignment, only the projection matrix is updated whereas the pre-trained model is freezed. This strategy is beneficial in reducing the computation costs and preventing the problem of memory-interference and memory loss [1]–[6].

II. Related Work

Conventional Methodology. Plant disease detection and classification is a classic research problem that has been studied extensively by researchers. Research scholars came up with many solutions that involved the use of some alterations, augmentations and various variants of models primarily from Machine Learning and Deep Learning like the Random Forests, Support Vector Machines, Convolutional Neural Networks etc [8], [11]–[16]. As understood, the above approaches were strictly limited to the problem statement and did not make extra efforts to explore further [17] which could possibly be the reason for their downfall.

Contemporary Methodology. The traditional models were replaced by the development of multimodal models [18] and transformers [19] that were based on the attention mechanism module which led to the production of better results. The rise of vision transformers [20] stepped up the game and produced better classification results than the transformers [21]–[23]. But even so, the models lacked interactivity with the user.

Visual Question Answering Models (VQA). The limitations posed by the above practices were solved by the evolution of the VQAs [24]. VQAs provide an interactive environment for the user to connect with the model and go beyond the go-to task of classification and detection. The model has the ability to answer free-form and open-ended questions asked by the user about the image and the user can expect the answer back in natural language. There have been several variations and advancements in the architecture of this multimodal model [25]. [26] conducts a detailed experiment to make a comparison and evaluate the performance of Stacked Attention Network Model (SANM), bidirectional Long Short Term Memory (LSTM) and Multimodal Tucker Fusion (MUTAN) and the results depict that VQA produces result with the best accuracy and minimal loss.

Large Language and Vision Assistant (LLaVA). LLaVA is an end-to-end trained large multimodal model that enables engaging dialogue based conversations on visual inputs [1]. The architecture conjoins visual encoder and Vicuna to generate a conversation or chat between the user and the model about the visual comprehension of the input.

Phytopathological-Specific Conversation. The engagement of the general-task based LLaVA was experimented on the phytopathological domain and it was found that the model demonstrated an inability to generate accurate responses, providing instead generalised solutions that exhibited universality across all instances of leaves.

III. Phytopathological Visual Question-Answer Data Synthesis

In plant disease diagnosis, sufficient image-annotation data pairs are accessible in the public domain, with the PlantVillage dataset being one of the most common and widely utilized resource. However, the lack of a multi-domain plant disease

TABLE I: Responses from Different Models

Fig. 2: Sample image of a strawberry leaf with scorch disease.

Q: Describe the condition/disease associated with the leaf in the image.

Model	Response
GPT-4 Vision	The leaf in the image appears to be infected with a type of fungal disease known as leaf spot. Leaf spot diseases cause the discoloration and blemishes seen here, and they can be caused by various fungal or sometimes bacterial pathogens. To prevent leaf spot diseases, you can take various measures: 1) Plant Disease-Resistant Varieties: Whenever possible, choose plant species and varieties that are known to be resistant to leaf spot and other diseases. 2) Proper Spacing: Ensure that plants have sufficient space between them to improve air circulation, which can help reduce the humidity that promotes the growth of fungi and bacteria.
LLaVA	The leaf is infected with a fungal disease, as indicated by the presence of spots and holes on its surface. To prevent the spread of this disease, it is essential to practice proper plant care and maintenance. This includes regular pruning, ensuring adequate air circulation, and maintaining a balanced soil pH. Additionally, using fungicides or organic treatments, such as neem oil or compost tea, can help protect the plant from fungal infections. It is also important to monitor the plant's health and remove any infected leaves or branches promptly to prevent the disease from spreading to other parts of the plant.
LLaVA-PlantDiag	The leaf in the image appears to be infected with leaf scorch disease. Leaf scorch is a common disease that affects strawberry plants and is characterized by browning and necrosis of the leaf tissue. It can be caused by various factors such as fungal or bacterial pathogens, environmental stress, or nutrient deficiencies. To prevent leaf scorch, it is important to maintain proper plant hygiene and avoid overhead irrigation. Watering at the base of the plant and providing adequate spacing between plants can help improve air circulation. Additionally, ensuring proper nutrition and avoiding stress factors such as extreme temperatures or excessive fertilization can also help prevent the disease.

diagnosis data limits the advancements of multimodal applications in this field. To address this limitation, the proposed research presents the generation of a novel phytopathological multimodal instruction-following data for visual instruction tuning, enabling the training of multimodal models in plant diagnosis.

A. Phytopathological Image Feature Alignment Data

The PlantVillage dataset contains 54,305 single leaf images from 14 different crops which are classified amongst 38 classes as species disease or species healthy [7]. To facilitate the preparation of data for concept alignment within the domain, a random selection of 200 images is made from each class, ensuring that the upper limit for classes with fewer images was reached. This selection process resulted in the development of a down-sampled dataset containing 6,324 images, each paired with an image X_v and its corresponding caption X_c. The caption in PlantVillage for an image was the label of the disease the plant leaf was infected with, i.e., Plant_LeafDisease.JPG. Utilizing a multimodal model, the captions associated with each image were elongated, forming extended captions, X_{ec}.

Therefore, for a phytopathological image X_v and its corresponding extended caption X_{ec} which is formulated from the caption X_c , we sample a question X_q, which asks to describe the phytopathological image. With (X_v, X_c, X_{ec}, X_q), we create a single-round instruction-following example:

Transition: $\mathbf{X_c} \rightarrow \mathbf{X_{ec}}$
Human: $\mathbf{X_q}$ $\mathbf{X_v}$ <STOP> ***Assistant:*** $\mathbf{X_{ec}}$ <STOP>

Given Image: `Potato_healthy_37.JPG` **`(X_v)`**
Formulation:

- **`X_c:`** `Potato_healthy_37`
- **`X_q:`** `How would you describe the overall appearance of the leaf in the image?`
- **`LLM generated X_ec:`** `The leaf in the image appears vibrant and healthy, with a full range of colors. There are no visible signs of disease or distress, indicating a robust and thriving potato plant.`

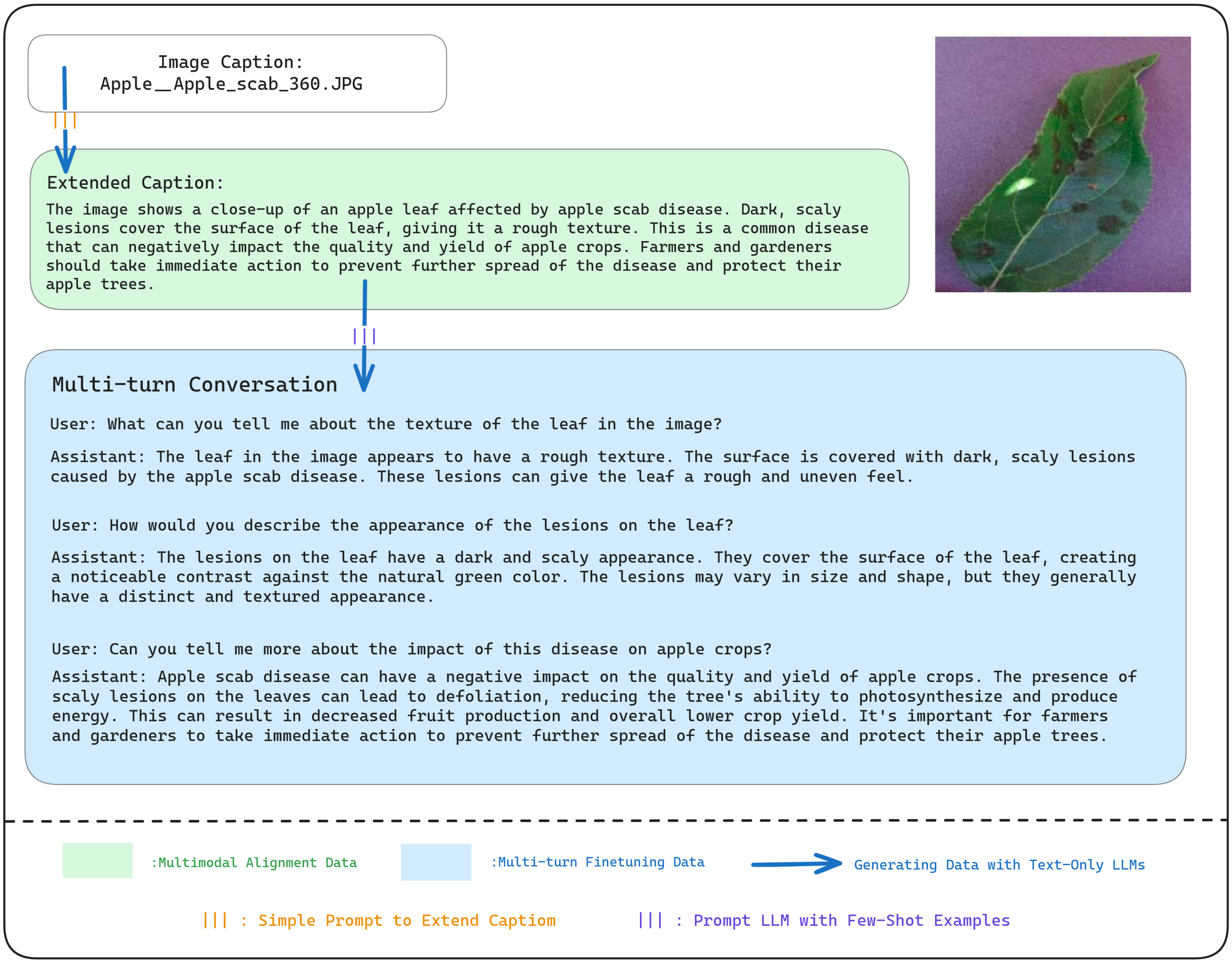

Fig. 3: An example of a multi-round conversation generated by GPT-3.5 showing the flow of data synthesis

The questions are drawn at random from the pool of concise, one-line questions .

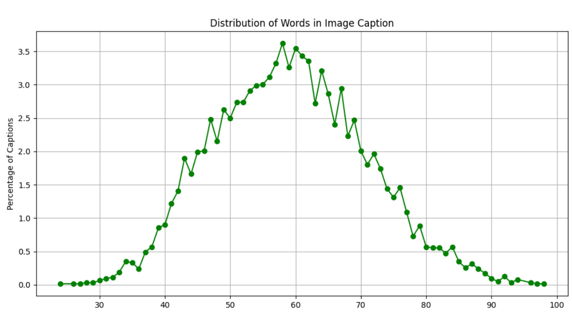

Fig. 4: Distribution of the caption length in Phytopathological Concept Alignment Data

B. Phytopathological Instruction-Tuning Data for Language Model

To enable the interactive assistant functionality within the language model, the multi-round conversations are initiated which are facilitated by the GPT-3.5. Leveraging context-alignment data, the iterative dialogues related to phytopathological images are generated without explicitly providing the images. Instead, the image information is conveyed through captions, prompting the language model, GPT-3.5, to engage in question-and-answer interactions, simulating an understanding as if it had visual perception. Using prompt engineering, the prompts are optimized for efficient instruction-following in the approach. Further ahead, a set of some few-shot examples in the prompts is manually selected and created to illustrate the process of generating quality conversations. These examples serve as practical demonstrations of effectively using captions and context for meaningful interactions. An illustration of the instruction-following data is presented in Fig 3 along with the data insights in Fig 6 and Fig 7.

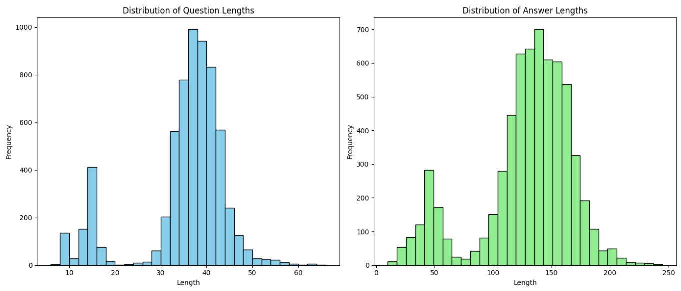

Fig. 5: Distribution of the length of question-answer pair produced in multi-round conversations

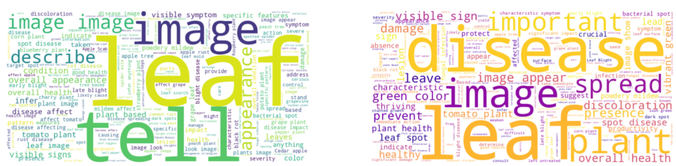

Fig. 6: Word Cloud representation of Questions (Left) and Responses (Right)

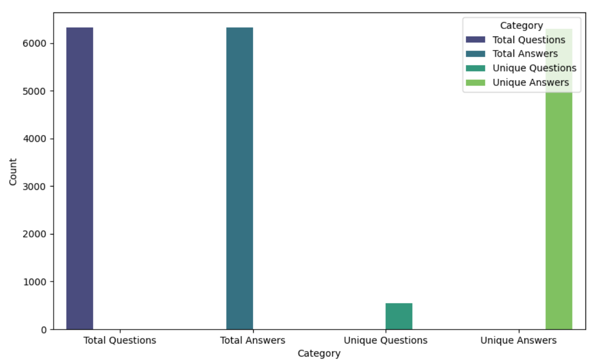

Fig. 7: Variation and Distribution of Questions and Answers

IV. Customizing LLaVA to Phytopathological Domain

To adapt the general-domain LLaVA model to the phytopathological domain, a two-step procedure was followed which involved pre-training followed by fine-tuning.

Model Architecture: LLaVA-PlantDiag is based on LLaVA-1.5 model. It consists of a large language model namely Vicuna-13B connected with pre-trained CLIP visual encoder ViT-L/14 using a two layer MLP and LoRA module. MLP layer is a trainable projection matrix/layer which maps the visual features from CLIP encoder to the word embedding space of the Vicuna-13B. These tokens share the same dimensionality as the word embedding space within the Vicuna-13B language model.

Pre-training LLaVA: In order to validate image feature alignment to textual embedding, the PlantVillage dataset was customised to the domain-specific needs. The images with short captions were extended and made more descriptive. With the help of the generated description, a structured conversation was formed using question-answer pair. In order to maximise the liklihood of trainable parameters, the weights of the visual encoder and LLM were kept frozen through out the training [1].

$$p\left(\mathbf{X}_{\mathrm{a}} \mid \mathbf{X}_{\mathrm{v}}, \mathbf{X}_{\text{instruct}}\right)=\prod_{i=1}^{L} p_{\boldsymbol{\theta}}\left(x_{i} \mid \mathbf{X}_{\mathrm{v}}, \mathbf{X}_{\text{instruct},<i}, \mathbf{X}_{\mathrm{a},<i}\right) \tag{1}$$

We compute probability of target X_a for sequence length L, given the image X_v and multi-turn conversation $X_{instructt}$ using equation 1

Fine-tuning LLaVA: To equip the model to carry out an insightful conversation on plant pathology, the model is fine-tuned on the custom phytopathological multimodal instruction-following dataset. The step involves the updation of the trainable parameters only, while the weights of visual encoder are kept frozen.

Fine-tuning LLaVA with LoRA: Due to the excessively large number of trainable parameters in LLaVa (around 13 billion parameters), conducting full-parameter fine-tuning can be very expensive and less practical. Hence, we opted for Low-Rank Adaptation (LoRA) fine-tuning. In LoRA, the pre-trained model weights are frozen, and trainable rank decomposition matrices are introduced. This significantly reduces the number of trainable parameters, making the training process more efficient. LoRA involves storing model weights in memory instead of updating them directly, a process referred to as ”freezing” the weights of the model. This approach is suitable for models like LLM, which is extensive and versatile, demonstrating proficiency across various tasks. For training in the phytopathological domain, where precise alignment of the model to the task is not crucial, a 128-rank LoRA fine-tuning is applied, reducing the trainable parameters from 13 billion to around 4 million (about 0.031% of trainable parameters among the total model parameters).

V. Experiments

A. Dataset Statistics

The phytopathological multimodal instruction-following data was constructed using 6,324 images from the PlantVillage dataset. The labels of the images were extended to produce extended captions, X_{ec}. The word length distribution of X_{ec} exhibits a unimodal, bell-shape like, symmetrical pattern, with the mean length containing 59.19 words in the extended captions. The following is represented in Fig4. The extended captions were used to make the question-answer dataset. The length distribution curve for the question-answers exhibits a bimodal pattern, with mean question length of 34.82 and mean answer length of 128.61. Fig 5 allows for the interpretation of the following insights.

B. Implementation Details:

The LLaVA-PlantDiag model is trained on a specialized dataset designed for identifying plant diseases and tested its performance on two key tasks: multi-turn conversation (i.e., question-answering) and disease classification. The foundational models for LLaVA-PlantDiag were Vicuna-1.3 and the CLIP ViT-L/14. The training utilized four, L40 GPUs, each with a memory capacity of 45GB. During pre-training, the model was trained for one epoch using Image Feature Alignment Data with a batch size of 64, a learning rate of 1e-3, and employing the AdamW optimizer, taking a total time of 38 minutes. During the fine-tuning phase, phytopathological instruction-tuning data was utilized. The model underwent training for another epoch, but with a reduced batch size of 32 and a lower learning rate of 2e-4. Due to the limited amount of task-specific data, there was an integration of the LoRA technique to adapt the model more effectively to our tasks. The fine-tuning stage was completed in 1 hour 48 minutes. The entire process aimed to equip LLaVA-PlantDiag with robust capabilities for both understanding queries in a conversational context as well as accurately classifying plant diseases.

C. Evaluation Metrics

1) Multi-round Conversations: To analyze the methodical insight of the multi-turn conversations generated by LLaVA-PlantDiag on the phytopathological multimodal instruction-following data, GPT-4 has been employed to asses the quality of generated responses. The contender models (e.g., LLaVA) are provided with an image and a question. The responses generated by the models are evaluated by comparing them with the descriptive text considered as the ground truth. In this assessment, GPT-4 functions as a judge, assigning scores to the models based on their performance.

2) Classification: To assess the classification performance of LLaVA-PlantDiag in contrast to other contender models, the evaluation utilizes the consideration of the accuracy metric to measure the classification effectiveness of these models.

D. Performance on Phytopathological Visual Question Answering Task

To evaluate our model, a test set of 250 questions and 90 images was created which was unseen by our model from the PlantVillage dataset. It was followed by a comparison of LLaVA-PlantDiag against various open-source and closed-source models like mPLUG-OWL, IDEFICUS, GPT-4 Vision and Gemini respectively. LLaVA-PlantDiag with LoRA as reported in TABLEII even outperforms GPT-4 Vision and Gemini which is trained on very large set of plant images from the internet. Furthermore, as indicated in TABLE I, both GPT-4 Vision and LLaVA failed to provide precise descriptions of the leaves, often giving broad, generic responses. In contrast, LLaVA-PlantDiag gave accurate, specific answers and also provided correct advice tailored to the condition of leaf.

TABLE II: Model Performance On Multi-turn Conversation

Model	**Score ↑**	**Input Modalities**	**LLM**
	Closed Source		
GPT-4 Vision	48.7	T, I	–
Gemini	42.4	T, I	–
	Open Source		
IDEFICUS	46.9	T, I	LLaMA-7B
mPlug-OWL	40.8	T, I	LLaMA-7B
LLaVA-PlantDiag (13B)	64.7	T, I	Vicuna-13B

TABLE III: Model Performance On Classification Tasks

Model	**Acc * ↑**	**Input Modalities**	**LLM**
	Closed Source		
GPT-4 Vision	24	I	–
Gemini	65	I	–
	Open Source		
IDEFICUS	85	I	LLaMA-7B
mPlug-OWL	23	I	LLaMA-7B
InstructBLIP	48	I	Vicuna-13B
LLaVA-PlantDiag (13B)	96	I	Vicuna-13B

* Acc. stands for accuracy

E. Performance on Phytopathological Classification Task

The classification performance of LLaVA-PlantDiag was assessed and compared with that of other models. A set of 100 images, previously unseen by our model, was randomly sampled from the PlantVillage dataset. The results, presented in TABLE III, highlight the significant performance of the LLaVA-PlantDiag as compared to other models, giving an important insight on the limited classification capability of other models on the multi-turn phytopathological visual question-answering task.

VI. Conclusion and Future Work

This paper proposes LLaVA-PlantDiag, a large language-and-vision model accustomed to the phytopathological domain. The model utilizes a unique VQA dataset, representing the phytopathological multimodal instruction-following data which is not yet explored by the researchers. LLaVA-PlantDiag exhibits better performance on Vision-Language as well as classification tasks when compared to other SoTA models in plant phytopathology. The upcoming research aims to enhance the dataset by including a wider range of examples and adjusting the model's structure to reduce any biases it might have, when dealing with images from similar categories.

References

[1] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[2] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models, 2022.

[3] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023.

[4] Bulent Tugrul, Elhoucine Elfatimi, and Recep Eryigit. Convolutional neural networks in detection of plant leaf diseases: A review. *Agriculture*, 12(8), 2022.

[5] Aydin Kaya, Ali Seydi Keçeli, Cagatay Catal, Hamdi Yalic, Temuçin Huseyin, and Bedir Tekinerdogan. Analysis of transfer learning for deep neural network based plant classification models. *Computers and Electronics in Agriculture*, 03 2019.

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[7] Mehmet Alican Noyan. Uncovering bias in the plantvillage dataset, 2022.

[8] Muhammad Hammad Saleem, Johan Potgieter, and Khalid Mahmood Arif. Plant disease detection and classification by deep learning. *Plants*, 8(11), 2019.

[9] Sachin D. Khirade and A.B. Patil. Plant disease detection using image processing. In *2015 International Conference on Computing Communication Control and Automation*, pages 768–771, 2015.

[10] Konstantinos P. Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, 2018.

[11] Murk Chohan, Adil Khan, Rozina Chohan, Saif Katper, and Muhammad Mahar. Plant disease detection using deep learning. *International Journal of Recent Technology and Engineering*, 9:909–914, 04 2020.

[12] Punam Bedi and Pushkar Gole. Plant disease detection using hybrid model based on convolutional autoencoder and convolutional neural network. *Artificial Intelligence in Agriculture*, 5:90–101, 2021.

[13] Arunabha M. Roy and Jayabrata Bhaduri. A deep learning enabled multi-class plant disease detection model based on computer vision. *AI*, 2(3):413–428, 2021.

[14] Shima Ramesh, Ramachandra Hebbar, Niveditha M., Pooja R., Prasad Bhat N., Shashank N., and Vinod P.V. Plant disease detection using machine learning. In *2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, pages 41–45, 2018.

[15] Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Mohamed Arselene Ayari, Aftab Ullah Khan, Muhammad Salman Khan, Nasser Al-Emadi, Mamun Bin Ibne Reaz, Mohammad Tariqul Islam, and Sawal Hamid Md Ali. Automatic and reliable leaf disease detection using deep learning techniques. *AgriEngineering*, 3(2):294–312, 2021.

[16] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 2016.

[17] Marko Arsenovic, Mirjana Karanovic, Srdjan Sladojevic, Andras Anderla, and Darko Stefanovic. Solving current limitations of deep learning based approaches for plant disease detection. *Symmetry*, 11(7), 2019.

[18] Md Shofiqul Islam, Sunjida Sultana, Fahmid Al Farid, Md Nahidul Islam, Mamunur Rashid, Bifta Sama Bari, Noramiza Hashim, and Mohd Nizam Husen. Multimodal hybrid deep learning approach to detect tomato leaf disease using attention based dilated convolution feature extractor with logistic regression classification. *Sensors*, 22(16), 2022.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[21] Rayene Amina Boukabouya, Abdelouahab Moussaoui, and Mohamed Berrimi. Vision transformer based models for plant disease detection and diagnosis. In *2022 5th International Symposium on Informatics and its Applications (ISIA)*, pages 1–6, 2022.

[22] Guoqiang Li, Yuchao Wang, Qing Zhao, Peiyan Yuan, and Baofang Chang. Pmvt: a lightweight vision transformer for plant disease identification on mobile devices. *Frontiers in Plant Science*, 14, 2023.

[23] Pushkar Gole, Punam Bedi, Sudeep Marwaha, Md. Ashraful Haque, and Chandan Kumar Deb. Trincnet: a lightweight vision transformer network for identification of plant diseases. *Frontiers in Plant Science*, 14, 2023.

[24] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[25] Venkat Kodali and Daniel Berleant. Recent, rapid advancement in visual question answering architecture: a review, 2022.

[26] Souvik Chowdhury and Badal Soni. eaVQA: An experimental analysis on visual question answering models. In Sivaji Bandyopadhyay, Sobha Lalitha Devi, and Pushpak Bhattacharyya, editors, *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 550–554, National Institute of Technology Silchar, Silchar, India, December 2021. NLP Association of India (NLPAI).