

# A Multimodal Benchmark Dataset and Model for Crop Disease Diagnosis

Xiang Liu<sup>†1,2</sup>, Zhaoxiang Liu<sup>†\*1,2</sup>, Huan Hu<sup>1,2</sup>, Zezhou Chen<sup>1,2</sup>, Kohou Wang<sup>1,2</sup>, Kai Wang<sup>1,2</sup>, and Shiguo Lian<sup>\*1,2</sup>

<sup>1</sup> AI Innovation Center, China Unicom, Beijing 100013, China

<sup>2</sup> Unicom Digital Technology, China Unicom, Beijing 100013, China

{liux750,liuwx178,huh30,chenzz51,wangzp103,wangk115,liansg}@chinaunicom.cn

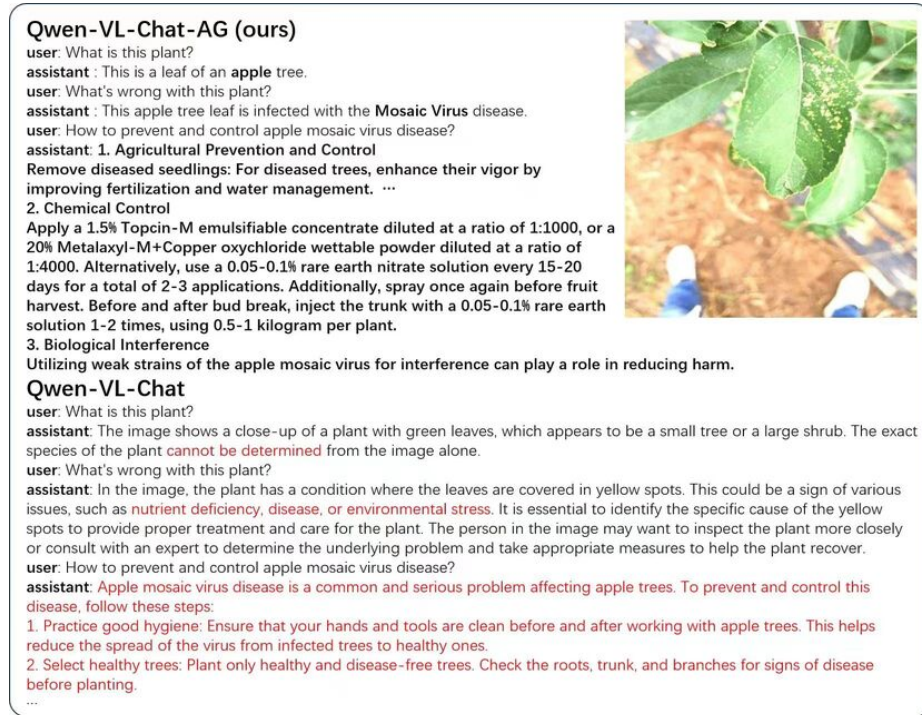
<sup>†</sup>Equal contribution, <sup>\*</sup>Corresponding author(s)

**Abstract.** While conversational generative AI has shown considerable potential in enhancing decision-making for agricultural professionals, its exploration has predominantly been anchored in text-based interactions. The evolution of multimodal conversational AI, leveraging vast amounts of image-text data from diverse sources, marks a significant stride forward. However, the application of such advanced vision-language models in the agricultural domain, particularly for crop disease diagnosis, remains underexplored. In this work, we present the crop disease domain multimodal (CDDM) dataset, a pioneering resource designed to advance the field of agricultural research through the application of multimodal learning techniques. The dataset comprises 137,000 images of various crop diseases, accompanied by 1 million question-answer pairs that span a broad spectrum of agricultural knowledge, from disease identification to management practices. By integrating visual and textual data, CDDM facilitates the development of sophisticated question-answering systems capable of providing precise, useful advice to farmers and agricultural professionals. We demonstrate the utility of the dataset by finetuning state-of-the-art multimodal models, showcasing significant improvements in crop disease diagnosis. Specifically, we employed a novel finetuning strategy that utilizes low-rank adaptation (LoRA) to finetune the visual encoder, adapter and language model simultaneously. Our contributions include not only the dataset but also a finetuning strategy and a benchmark to stimulate further research in agricultural technology, aiming to bridge the gap between advanced AI techniques and practical agricultural applications. The dataset is available at <https://github.com/UnicomAI/UnicomBenchmark/tree/main/CDDMBench>.

**Keywords:** Multimodal dataset · Crop disease diagnosis · Large-scale vision-language model

## 1 Introduction

The accurate diagnosis of crop diseases is pivotal in agricultural practices, offering invaluable support to individuals involved in farming activities. However,



**Fig. 1:** Example comparison of LVLMs on crop disease diagnosis. Our model accurately identifies crop and disease categories, offering detailed prevention and treatment methods. In contrast, Qwen-VL-Chat fails to determine both crop and disease categories, and provides detailed prevention and treatment methods, as indicated by the red texts.

prevailing solutions for crop disease diagnosis predominantly rely on single-modal approaches, such as visual algorithms for classification [2, 18] and detection [28]. These single-modal methods can only provide diagnosis results and cannot offer richer agricultural knowledge based on human preferences. The imperative for a multimodal visual question answering (VQA) system [5, 13], proficient in diagnosing crop diseases while possessing extensive agricultural knowledge, becomes evident. Such a system, facilitated by multimodal interactions and multi-round conversations, has the potential to expedite the retrieval of information concerning specific diseases and recommend appropriate mitigation strategies.

Most popular large-scale vision-language models (LVLMs), such as minigpt4 [30], Flamingo [3], LLaVA [24] and Qwen-VL [9], while effective as general-purpose multimodal conversational assistants [7, 14, 21], struggle in the crop disease domain. For example, Qwen-VL struggles with accurately identifying crop types, diagnosing diseases, and falls short in providing detailed strategies for disease prevention and control, as shown in Figure 1. Consequently, developing a multimodal dataset specifically for adapting LVLMs to the crop disease domain



**Fig. 2:** Examples of the crop disease image dataset. Each image represents a different category, and the leaves show a high degree of similarity, from their colors to their shapes. Additionally, some spot diseases display very similar visual features. Among the images, the two marked with red boxes represent different diseases but look very similar; the two marked with yellow boxes belong to different types of crops but have a very similar shape.

is essential to enhance their accuracy and utility in agriculture. Building on this need, our work introduces a CDDM dataset.

Current finetuning strategy for LVLs involves freezing the visual encoder while adjusting the projection/adaptor module and the language model [24]. However, this approach faces challenges in the crop disease domain, due to the similarity of different crop diseases (as shown in Figure 2). Employing this strategy means the visual encoder’s ability to differentiate similar samples is limited, ultimately impacting the accuracy of disease diagnosis.

The contributions of our paper can be summarized as follows:

- The CDDM Dataset: We have meticulously curated a comprehensive dataset, comprising 137,000 images of crop diseases, and constructed a diverse set of 1 million question-answering instances. As illustrated in Figure 3, for each diseased crop image, we’ve crafted conversations that encompass a range of information, including the crop and disease categories, detailed disease insights, and prevention and control strategies. This dataset serves as a fundamental resource for training models, which are capable of comprehending and addressing queries related to crop diseases.
- Model Finetuning Strategy: We adopted a novel finetuning strategy utilizing the LoRA [17], for training Qwen-VL-Chat on the CDDM dataset without

freezing the visual encoder. Our experiments have validated the effectiveness of this strategy, demonstrating its potential to significantly improve diagnostic accuracy in the agricultural domain.

- Open-Source Initiative: To promote research in the field of agricultural multimodal learning, we intend to release our dataset and the corresponding model codebase to the public. This initiative aims to foster collaboration and propel further advancements in the development of multimodal question-answering systems tailored for agricultural applications.

In conclusion, our work not only addresses the current limitations in crop disease diagnosis but also lays the groundwork for the development of specialized multimodal question-answering systems in the agricultural domain. Through the dissemination of our dataset and finetuned models, we aspire to pave the way for more effective and knowledgeable agricultural assistants that can significantly contribute to the enhancement of crop yield and overall farm management.

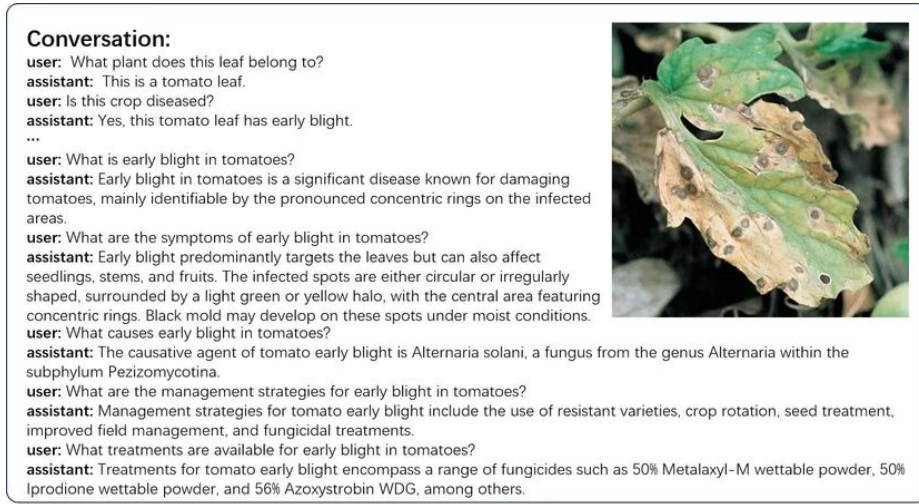
## 2 Related Work

### 2.1 Related Work on Traditional Agricultural Diseases Diagnosis

Traditional agriculture heavily relies on the observation and expertise of farmers and experts for the diagnosis and treatment of crop diseases. However, with the advent of smart agriculture, there is a growing trend towards utilizing computer vision methods to assist in disease diagnosis. Arya [6] employed AlexNet [19] architectures to detect diseases in mango and potato leaves. Yang [29] proposed a multimodal feature fusion network that combines RGB image networks with hyperspectral band extraction networks to improve the recognition accuracy of citrus Huanglongbing. Furthermore, Morbekar [25] utilized YOLO [27] for object detection to identify plant diseases, while Divyanth [12] introduced a two-stage segmentation method using YOLO for plant segmentation and a disease segmentation model. These methods offer limited information, falling short in effectively aiding farmers with crop disease prevention and control.

### 2.2 Multimodal Methods

Multimodal question answering combines image and text information for answering queries. In a conventional VQA model, an image and a question are input, with visual and language encoders embedding them respectively. Following this, a cross-modal attention block utilizes the embeddings from both visual and language sources to understand the connections between the image and the question, which is crucial for making the final prediction. Lan [20] created a question-answer pairs dataset for fruit tree diseases, employing a co-attention architecture aligning ResNet [16] and BERT [11] features to achieve simple agricultural multimodal question-answering. However, this VQA model supports a limited range of crop diseases. Due to the limitations of its architecture, it has



**Fig. 3:** An instance of our CDDM data. The conversations cover the diagnosis, prevention, and treatment of crop diseases.

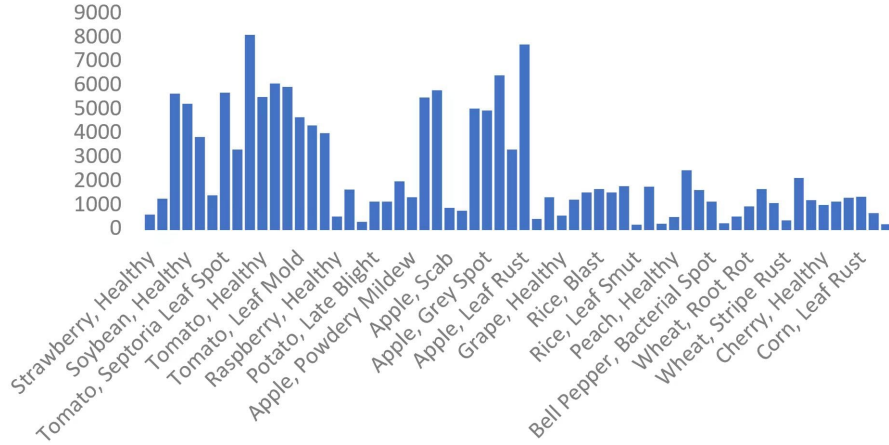
restricted generalization capabilities, offering only simplistic answers and lacking comprehensive agricultural knowledge.

The most popular LVLM typically combines a visual encoder and a large language model [1, 4, 8, 10, 15] via an adapter or a projector, where the former provides high-quality visual representation, and the latter offers powerful language generation capability. Li [23] proposed using a lightweight querying transformer to align image encoders and frozen large language models. Liu [24] introduced LLaVA model which is trained with multimodal instruction-following data [26], and achieves a relative score of 85.1% compared to GPT-4. Bai [9] presented the Qwen-VL model, which is trained on 1.14 billion image-text pairs. While these models excel in general domains, they are less effective in crop disease diagnosis because the image-text pairs specific to crop diseases differ significantly from the counterpart found on the web.

### 3 Crop Disease Domain Multimodal Dataset

To construct a robust crop disease diagnosis multimodal question-answering model, we developed a CDDM dataset. As depicted in Figure 3, we present a sample of the CDDM data, comprising an image of crops alongside a series of interactive conversations. This dialogue encompasses a range of information, including the identification of the crop, diagnosis of any diseases affecting it, as well as details regarding the causes of these diseases and measures for their prevention and control.





**Fig. 4:** Distribution of the number of images for crop diseases dataset.

Here, we outline the construction process from three key aspects: crop image data collection and annotation, crop disease diagnosis instruction-following data, and crop disease knowledge instruction-following data.

### 3.1 Crop Image Data Collection and Annotation

The image data for this study consists of two parts:

- Web Data: This includes agricultural datasets from Kaggle as well as agricultural disease data collected using web crawling methods, totaling 62,000 images.
- Private Data: We conducted field surveys in multiple farms and orchards to collect plant disease images, resulting in a total of 75,000 original images.

With the assistance of agricultural experts, we annotated all image data. The annotation information primarily includes crop category, disease category, and appearance description of the image.

Through data collection and annotation, we have compiled a dataset comprising 137,000 images. The dataset encompasses 16 categories of crops and a total of 60 categories of crop diseases. As illustrated in Figure 4, there are 48 categories each containing more than 500 images, while 7 categories feature image counts ranging from 200 to 500. Overall, the distribution of data across categories is relatively even.

In addition, we collected text content of crop diseases knowledge related to crop diseases covering detailed descriptions including crop, disease, disease symptoms, pathogen characteristics, transmission pathways, disease conditions, prevention and control methods.

**Prompt:**

As an AI visual assistant specialized in diagnosing crop diseases, I am examining a single image. What I see are three terms, for example, {Tomato, Early Blight, Description}. From three aspects (1. what type of plant leaf it is, 2. what disease it has, 3. the appearance characteristics of plant diseases), I'll describe the same image I'm observing. Responding in a manner as if seeing the image and answering questions, I'll design a dialogue between you and someone inquiring about this photo. Answers will reflect the tone of a visual AI assistant specialized in crop disease diagnosis. I'll present various questions and provide corresponding answers, including questions about visual content and crop disease-related inquiries. Only questions with definitive answers will be included: (1) those where content asked about can be confidently identified in the image, (2) those where it can be confidently determined that the content is not in the image, and (3) those related to disease information where a confident response can be provided. I will not ask any questions I cannot confidently answer. Also, complex questions related to image content, such as asking for background knowledge about objects in the image or discussing events in the image, will be included. Similarly, questions about uncertain details will be avoided. When answering complex questions, detailed responses will be provided, including examples or reasoning steps to enhance persuasiveness and organization. If necessary, multiple paragraphs will be included.

Example:

Input:

{Tomato, Early Blight, The leaves have yellow and brown patches in the middle.}

Output:

1. user: Describe the content of this image.

assistant: This image depicts an unhealthy tomato leaf with yellow and brown patches in the middle, indicating Early Blight.

2. user: What plant's leaf is this?

assistant: This is a tomato leaf.

3. user: Is this leaf from a cucumber plant?

assistant: No, this is a tomato leaf.

4. user: Is this a tomato leaf or a potato leaf?

assistant: This is a tomato leaf.

5. user: What disease does this tomato leaf have?

assistant: This tomato leaf is affected by Early Blight.

6. user: Is this crop diseased?

assistant: Yes, this tomato leaf is afflicted with Early Blight.

7. user: Has this leaf contracted Yellow Leaf Curl Virus?

assistant: No, this tomato leaf has Early Blight.

8. user: Is this apple leaf healthy?

assistant: This is not an apple leaf; it's a tomato leaf afflicted with Early Blight.

**Query:**

Input:

{Apple, Cedar Apple Rust, The leaf have two distinct lesions. The lesions are circular, with a black central spot surrounded by yellow and red ring-like areas.}

**GPT-4 output:**

1. user: Describe the content of this image.

assistant: This image depicts an unhealthy apple leaf with two distinct lesions. The lesions are circular, featuring a black central spot surrounded by yellow and red ring-like areas, indicative of Cedar Apple Rust.

2. user: What plant's leaf is this?

assistant: This is an apple leaf.

3. user: Is this leaf from an orange tree?

assistant: No, this is an apple leaf.

4. user: Is this an apple leaf or a pear leaf?

assistant: This is an apple leaf.

5. user: What disease does this apple leaf have?

assistant: This apple leaf is affected by Cedar Apple Rust.

6. user: Is this crop diseased?

assistant: Yes, this apple leaf is afflicted with Cedar Apple Rust.

7. user: Has this leaf contracted Apple Scab?

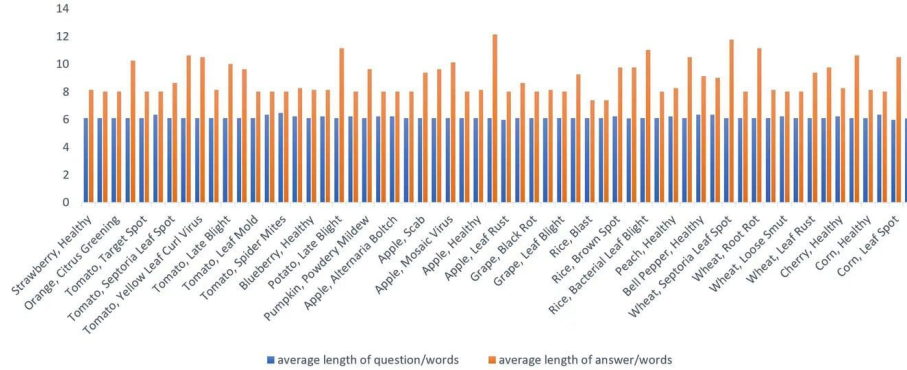
assistant: No, this apple leaf has Cedar Apple Rust.

8. user: Is this peach leaf healthy?

assistant: This is not a peach leaf; it's an apple leaf afflicted with Cedar Apple Rust.



**Fig. 5:** The prompt example of utilizing GPT-4 to generate instruction-following data of crop disease diagnosis. In the few-shot example within the "Prompt" part, the QA pairs highlighted in red are carefully crafted to include negative responses. After sequentially entering the "Prompt" part and the "Query" part, GPT-4 can generate 8 similar QA pairs, with negative responses highlighted in green.



**Fig. 6:** Distribution of the average lengths of questions and answers for different crop diseases in the CDDM Dataset.

### 3.2 Crop Disease Diagnosis Instruction-Following Data

To enable the model to accurately diagnose crop diseases, we generate diverse instruction-following data through multi-round conversations about the provided crop images, utilizing language-only GPT-4 prompting. Specifically, we design few-shot instructions in a prompt that asks GPT-4 to generate questions and answers, with input {crop category, disease category, appearance description}.

Experiments revealed that LVLMs tend to give affirmative responses more often in diagnosing plant species and disease categories. When posed with questions requiring negative answers, the visual model frequently errs by providing incorrect affirmative responses. Consequently, in crafting the question-answer corpus, we incorporated questions necessitating negative answers.

Figure 5 illustrates a prompt example for generating instruction-following data of crop disease diagnosis using GPT-4. In the "Prompt" part, we provide a few-shot example consisting of 8 questions and answers. The pairs highlighted in red are carefully crafted QA pairs with negative responses. After sequentially entering the content from the "Prompt" part and the "Query" part, GPT-4 can generate 8 similar QA pairs. The parts highlighted in green are the generated QA pairs with negative responses.

The instruction-following dataset of crop disease diagnosis contains over 1 million question-answer pairs. The average question length is 6.11 words, while the average answer length is 8.92 words. Figure 6 shows the distribution of the average lengths of questions and answers for 60 types of crop diseases.

### 3.3 Crop Disease Knowledge Instruction-Following Data

Based on the text content of crop disease knowledge we collected, we utilized GPT-4 to generate instruction-following conversations. The full prompt and the detailed example can be found in Figure 7. Below is an example from the content in the "Query" section:



**Prompt:**  
 You are a specialist in text paraphrasing, tasked with processing a given text in JSON format to generate a series of questions and answers. The JSON format like this  
 {\"title\":\"content\":\"content\"}\\nRequirements: \\n1. Your designed series of questions should encompass the topics highlighted in the \"title,\" while also delving beyond these confines. It's advisable to thoroughly analyze the \"content\" before formulating your questions. \\n2Your designed series of answers should encompass the topics highlighted in the \"content.\". \\n3. The desired output format is as follows: [{\"user\":\"question\",\"assistant\":\"answer\"},{\"user\":\"question\",\"assistant\":\"answer\"}...], without including any extraneous symbols. \\n4. Each question aligns with the thought process typical of the general population. \\n\\nInput JSON content:

**Query:**

```
{
  "title": "Tomato Early Blight",
  "summary": "Tomato Early Blight, also called 'ring pattern disease', is a common and major pest of tomatoes. ...",
  "sections": [
    {
      "title": "Tomato Early Blight , Harmful Symptoms",
      "content": "It primarily affects the leaves but can also damage seedlings, stems, and fruits. Infected seedlings show dark brown lesions at the base of the stem, slightly sunken with concentric rings. ..."
    },
    {
      "title": "Tomato Early Blight , Pathogens and Infection",
      "content": "Early blight is caused by the fungus Alternaria solani, belonging to the subdivision Deuteromycotina, genus Alternaria. Besides affecting tomatoes, it can also infect potatoes, eggplants, and peppers. The primary mode of infection is through conidia. ..."
    },
    {
      "title": "Tomato Early Blight , Prevention and Control Techniques",
      "content": "1. Choice of Variety\\nGenerally, early-maturing varieties and narrow-leaf varieties have lighter disease incidence, while tall, large seedling, and large-leaf varieties have heavier disease incidence.\\n..."
    }
  ]
}
```

**GPT-4 output:**

```
[
  {
    "user": "What is Tomato Early Blight and why is it a concern?",
    "assistant": "Tomato Early Blight, also known as 'ring pattern disease', is a common and significant pest problem for tomatoes, affecting a broad range of plant hosts including crops in the Solanaceae family such as eggplants, peppers, and potatoes. It's a concern because it leads to severe outbreaks, especially in localities that have shifted to growing more virus-resistant tomato varieties instead of those resistant to early blight, resulting in substantial crop damage."
  },
  {
    "user": "What environmental conditions favor the development of Tomato Early Blight?",
    "assistant": "The development of Tomato Early Blight is favored by higher temperatures and humidity. Conidia can germinate quickly in water droplets at temperatures of 28-30°C. Early summer conditions like frequent rains and fogs facilitate the quick and large-scale formation of conidia, making the disease easily spread. The disease's occurrence is closely related to the growth stage of the host, with the disease starting on the older leaves and rapidly developing as the plant ages."
  },
  {
    "user": "How can Tomato Early Blight be prevented and controlled?",
    "assistant": "Prevention and control of Tomato Early Blight involve a combination of strategies, including choosing early-maturing or narrow-leaf varieties, practicing crop rotation with non-solanaceous crops for more than two years, seed treatment using heat and warm water to eliminate pathogens, and chemical control. Chemical control involves cultivating strong seedlings, practicing high ridge cultivation, rational fertilization, enhancing greenhouse ventilation, and regular removal of diseased leaves. Starting prevention with fungicides like 'Frostbell' from the seedling stage is also recommended, alongside treatments for disease outbreaks."
  }
]
```

**Fig. 7:** The prompt example of utilizing GPT-4 to generate instruction-following data of crop disease knowledge. After sequentially entering the "Prompt" section and the "Query" section, GPT-4 can generate high-quality QA pairs.

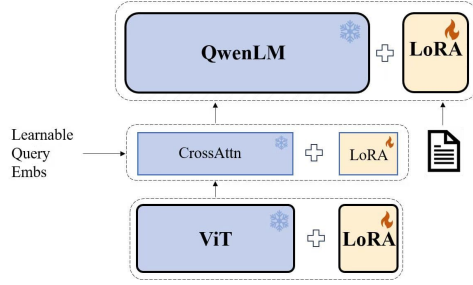
*{\"title\": \"Tomato Early Blight\", \"summary\": \"Tomato Early Blight, also called 'ring pattern disease', is a common and major pest of tomatoes. Some localities have been encouraged to grow more virus-resistant tomato varieties instead of those resistant to early blight, leading to severe blight outbreaks. The disease affects a broad range of plant hosts, including, but not limited to, crops in the Solanaceae family, such as eggplants, peppers, and potatoes.\"}*

Corresponding to this, the content of the question and answer pair generated by GPT-4 is presented as follows:

*{\"user\": \"What is Tomato Early Blight?\", \"assistant\": \"Tomato Early Blight, also known as 'ring pattern disease', is a common and major pest of tomatoes that affects a broad range of plant hosts, including crops in the Solanaceae family such as eggplants, peppers, and potatoes.\"}*

In the crop disease knowledge instruction-following dataset, the average question length is 9.69 words, while the average answer length is 130.41 words.

**The CDDM Dataset:** We constructed the CDDM dataset by merging crop disease diagnosis and knowledge instruction-following data and crop dis-



**Fig. 8:** The LoRA training strategy on Qwen-VL-Chat.

ease knowledge instruction-following data according to plant species and disease types.

## 4 Adapting Multimodal Conversational Models to the Crop Disease Domain

We introduce a novel approach capable of adapting general LVLMs to agriculture focused LVLMs. Here, we use Qwen-VL-Chat as an example. Qwen-VL-Chat integrates three critical components: a language model, a visual encoder, and a position-aware vision-language adapter, i.e. cross attention module. As illustrated in Figure 8, our method involves a specialized training regime, utilizing the LoRA technique, tailored specifically towards the domain of crop disease diagnosis.

Our primary goal is adapting the Qwen-VL-Chat model to accurately diagnose crop diseases. To achieve this, We employ the LoRA training technique to simultaneously adjust the parameters of all three components of the model: the language model, the visual encoder, and the position-aware vision-language adapter. Our finetuning strategy is quite different from the finetuning strategy in LLaVA and Qwen-VL-Chat where the parameters of the visual encoder are not updated during finetuning.

## 5 Experiments

To evaluate the utility of our proposed dataset and fine-tuning strategy, we fine-tuned the LLaVA model and the Qwen-VL-Chat model using two different strategies on our dataset. One strategy is without freezing the visual encoder, the other is with freezing the visual encoder.

The model versions and hyper-parameters are below:

**LLaVA-v1.5-7B** hyper-parameters: {batch size: 128, learning rate: 2e-4, epochs: 5, maximum sequence length: 2048, weight decay: 0}

**Qwen-VL-Chat-7B** hyper-parameters: {batch size: 128, learning rate: 1e-5, epochs: 5, maximum sequence length: 2048, weight decay: 0.1}

### 5.1 Evaluation Metrics

**Crop Disease Diagnosis Performance.** To assess the model’s efficacy in diagnosing crop diseases, we constructed a test set using 3,000 images that were not included in the training set. The test set included a variety of questions and answers to evaluate the model’s performance comprehensively. The performance was measured based on the accuracy of the model’s responses, specifically detecting the keywords of crop category and disease category in its answers.

**Crop Disease Knowledge VQA Performance.** Mirroring the evaluation methodology used by LLaVA and LLaVA-Med [22], we employed GPT-4 to assess the quality of model’s generated responses to crop disease knowledge questions. We started by manually selecting original data pertaining to 10 types of crop diseases and crafting 20 questions. We randomly selected 20 images in the 10 types of crop diseases. Responses were then solicited from candidate models based on the images and questions provided. To provide an approximate theoretical upper bound, We create a reference prediction based on the question and the text content of crop disease knowledge, using the text-only GPT-4.

The responses from both the candidate models and the reference GPT-4 predictions were evaluated by GPT-4 for their helpfulness, relevance, and accuracy. Each response was scored on a scale from 1 to 10, with higher scores indicating superior overall performance. We calculated the total score for the model across all questions, with the maximum possible score being 200. To normalize this to a 100-point scale, we converted the total score accordingly.

**Table 1:** Results on crop disease diagnosis and knowledge QA. \* indicates freezing the visual encoder.

Model	Crop Disease Diagnosis		Crop Disease Knowledge QA
	Crop Classification	Disease Classification	
Qwen-VL-Chat	28.4%	5.0%	41
Qwen-VL-Chat-AG*	84.4%	66.1%	88.5
Qwen-VL-Chat-AG	97.4%	91.5%	84
LLaVA-v1.5-7b	24.5%	5.9%	47.5
LLaVA-AG*	94.3%	82.1%	<b>98</b>
LLaVA-AG	<b>98.0%</b>	<b>91.8%</b>	96.5

### 5.2 Results

Table 1 shows the experimental results. As can be seen, the models finetuned on our dataset outperform the LLaVA model and Qwen-VL-Chat model with a large margin. Our dataset builds a connection between images and the concepts of crops and diseases. This connection helps the finetuned models align image features with the LLM word embedding and then improve the performances.

The base models have poor accuracies because they are unable to fully model the connection without training on our dataset.

The models finetuned without the visual encoder frozen outperform those finetuned with the visual encoder frozen considerably. Due to the similarity in appearance of different crops and different diseases (as shown in Figure 2), the frozen visual encoder falls short of capturing the local details and patterns that distinguish them because it is trained on the general domain dataset. Finetuning it on our dataset enhances its ability to capture these local details and patterns, which is why finetuning results in a significant performance jump.

In Figure 1, we showcase examples of dialogues between Qwen-VL-Chat-AG and Qwen-VL-Chat in the context of crop disease diagnosis. Qwen-VL-Chat-AG was able to precisely identify crop diseases and offer effective prevention and treatment solutions, demonstrating the significant value of the CDDM dataset and the finetuning strategy in developing professional agricultural chatbots.

### 5.3 Limitations

**Handling diseases out of domain:** We conducted several tests and found that our fine-tuned models perform not well in handling diseases outside our dataset. We guess that in-context learning might be the potential solution to this issue, e.g., adding one/few examples out of domain in the prompt to guide the models handling diseases in and out of domains. And we leave it as our future work to explore.

## 6 Conclusions

We presented a CDDM dataset and a LoRA based finetuning strategy. A series of experiments are conducted to validate the utility of our dataset and the finetuning strategy. The models trained on our dataset with the proposed finetuning strategy gain significantly in the performances of crop disease diagnosis and knowledge VQA. Our contributions include not only the dataset but also a finetuning strategy and a benchmark to stimulate further research in agricultural technology, aiming to bridge the gap between advanced AI techniques and practical agricultural applications.

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Agarwal, M., Sinha, A., Gupta, S.K., Mishra, D., Mishra, R.: Potato crop disease classification using convolutional neural network. In: Smart Systems and IoT: Innovations in Computing: Proceeding of SSIC 2019. pp. 391–400. Springer (2020)

3. 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. *Advances in Neural Information Processing Systems* **35**, 23716–23736 (2022)
4. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023)
5. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2425–2433 (2015)
6. Arya, S., Singh, R.: A comparative study of cnn and alexnet for detection of disease in potato and mango leaf. In: *2019 International conference on issues and challenges in intelligent computing techniques (ICICT)*. vol. 1, pp. 1–6. IEEE (2019)
7. Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al.: A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* (2021)
8. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023)
9. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023)
10. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners advances in neural information processing systems 33 (2020)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018)
12. Divyanth, L., Ahmad, A., Saraswat, D.: A two-stage deep-learning based segmentation model for crop disease quantification based on corn field imagery. *Smart Agricultural Technology* **3**, 100108 (2023)
13. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multi-modal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016)
14. Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., Gao, J., et al.: Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision* **14**(3–4), 163–352 (2022)
15. 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)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021)
18. Khamparia, A., Saini, G., Gupta, D., Khanna, A., Tiwari, S., de Albuquerque, V.H.C.: Seasonal crops disease prediction and classification using deep convolutional encoder network. *Circuits, Systems, and Signal Processing* **39**, 818–836 (2020)



19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012)
20. Lan, Y., Guo, Y., Chen, Q., Lin, S., Chen, Y., Deng, X.: Visual question answering model for fruit tree disease decision-making based on multimodal deep learning. *Frontiers in Plant Science* **13**, 1064399 (2023)
21. Li, C., Liu, H., Li, L., Zhang, P., Aneja, J., Yang, J., Jin, P., Hu, H., Liu, Z., Lee, Y.J., et al.: Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems* **35**, 9287–9301 (2022)
22. 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. *Advances in Neural Information Processing Systems* **36** (2024)
23. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023)
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36** (2024)
25. Morbekar, A., Parihar, A., Jadhav, R.: Crop disease detection using yolo. In: 2020 international conference for emerging technology (INCET). pp. 1–5. IEEE (2020)
26. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277* (2023)
27. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 779–788 (2016)
28. Saleem, M.H., Potgieter, J., Arif, K.M.: Plant disease detection and classification by deep learning. *Plants* **8**(11), 468 (2019)
29. Yang, D., Wang, F., Hu, Y., Lan, Y., Deng, X.: Citrus huanglongbing detection based on multi-modal feature fusion learning. *Frontiers in Plant Science* **12**, 809506 (2021)
30. 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)