




OPEN

A large language model for multimodal identification of crop diseases and pests

Yiqun Wang¹, Fahai Wang^{1,3}, Wenbai Chen¹ [✉], Bowen Lv^{1,3}, Mengchen Liu^{1,3}, Xiangyuan Kong¹, Chunjiang Zhao² & Zhaocen Pan¹

Pests and diseases significantly impact the growth and development of crops. When attempting to precisely identify disease characteristics in crop images through dialogue, existing multimodal models face numerous challenges, often leading to misinterpretation and incorrect feedback regarding disease information. This paper proposed a large language model for multimodal identification of crop diseases and pests, which can be called LLMI-CDP. It builds up on the VisualGLM model and introduces improvements to achieve precise identification of agricultural crop disease and pest images, along with providing professional recommendations for relevant preventive measures. The use of Low-Rank Adaptation (LoRA) technology, which adjusts the weights of pre-trained models, achieves significant performance improvements with a minimal increase in parameters. This ensures the precise capture and efficient identification of crop pest and disease characteristics, greatly enhancing the model's application flexibility and accuracy in the field of pest and disease recognition. Simultaneously, the model incorporates the Q-Former framework for effective modal alignment between language models and image features. Through this approach, the LLMI-CDP model is able to more deeply understand and process the complex relationships between language and visual information, further enhancing its performance in multimodal recognition tasks. Experiments are carried out in the homemade datasets, The results demonstrate that the LLMI-CDP model surpasses five leading multimodal large language models in relevant evaluation metrics, confirming its outstanding performance in Chinese multimodal dialogues related to agriculture.

Keywords Large language model, Crop disease identification, Agricultural questions and answers, Multimodal

The recognition of agricultural crop disease and pest images along with knowledge-based question answering represents an essential feature in driving the development of smart agriculture¹. Currently, research on large language models (LLMs) for multimodal tasks such as question answering remains relatively scarce, especially in the domain of agricultural crop disease and pests' recognition and prevention. LLMs demonstrate strong reasoning capabilities across various domains². Models such as ChatGPT³, LLaMA⁴, GPT-4⁵, ChatGLM⁶, and PaLM⁷ have demonstrated remarkable capabilities in downstream tasks. These models achieve a profound understanding of the grammatical and semantic properties of natural language by undergoing extensive training on a broad spectrum of textual corpora. Their remarkable abilities, such as language comprehension, reasoning, and language generation, demonstrate their potential for widespread use in various fields. The current advantage of LLMs primarily lies in their proficiency in open-domain knowledge. Directly generating answers for vertical domains often fails to meet professional standards. Nevertheless, the potential natural language understanding abilities learned by these large models from general domains can be applied to other linguistic tasks. ChatLaw⁸, based on Ziya-LLaMA-13B, creates a legal language model by fine-tuning it with legal data and integrating vector database retrieval. DoctorGLM⁹, constructed on ChatGLM-6B and fine-tuned using a Chinese medical dialogue dataset, forms a model for Chinese medical consultation. BenTsao¹⁰, built on LLaMA-7B, utilizes a medical knowledge graph and GPT-3.5 API to create a dataset for Traditional Chinese Medicine (TCM) teaching, developing a TCM language model. Cornucopia¹¹, also based on LLaMA-7B, constructs a command dataset using publicly available Chinese financial data and scraped financial information, focusing on question-answering in the financial domain.

¹School of Automation, Beijing Information Science and Technology University, Beijing 100192, China. ²National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China. ³Beijing Research Center for Information Technology in Agriculture, Beijing 100097, China. ✉email: chenwb@bistu.edu.cn

Deep learning has demonstrated exceptional performance in the domain of image recognition and has emerged as a pivotal driver in advancing technological innovations for agricultural disease detection¹². Currently, a variety of deep learning-based convolutional neural networks (CNNs) and MobileNetV2 models have been successfully employed in the recognition of crop diseases¹³, yielding noteworthy results. However, it is important to note that these models are typically limited to the recognition and classification of disease-related images pertaining to individual crops¹⁴, without providing a comprehensive description of the disease's specific characteristics. While these models contribute positively to crop prevention efforts, they fall short in delivering timely and actionable control measures to farmers. A singular model, by its nature, is insufficient to meet the complex requirements of integrated disease management. Therefore, the development of a large-scale visual-language model could serve to unify disease recognition with decision-making processes for control measures, thereby enhancing the efficacy of crop disease prevention and management.

The increasing scope of LLMs, particularly in the multimodal space, and their careful application, represent a highly relevant area of study. Before the advent of widespread multimodal models, common approaches involved manually annotating images and videos¹⁵. Despite this, as the datasets grew larger and more complex, reliance solely on manual annotation proved insufficient to meet the demands for data quality and annotation efficiency. Advancements in multimodal technology have facilitated the transformation of heterogeneous multimodal data across images and videos into textual representations by extending models¹⁶. This shift aids in a more comprehensive and efficient assimilation of information. In traditional agriculture, the diagnosis and decision-making related to crop diseases and pests rely heavily on the observations of experts or experienced farmers, who promptly remove diseased plants. Conversely, in practice, diagnosing diseased plants first involves identification by experts, followed by consulting agricultural knowledge to determine treatment methods. The general trend of fine-tuning models for specialized vertical domains has become inevitable, driving the long-term development of artificial intelligence scenarios.

Existing large multimodal models struggle to accurately identify crop diseases and pests, and their classification performance for different diseases within the same crop is subpar. The absence of a specialized agricultural knowledge base limits these models to providing only general responses during question-and-answer sessions, which fail to offer effective guidance for pest and disease management. This paper introduces the large language model for multimodal identification of crop diseases and pests (LLMI-CDP). The primary objective of this model is to extend the capabilities of the large language model ChatGLM, which is initially designed for text-based question answering, into a multimodal architecture. This extension empowers the language model with consistent generation abilities across both language and visual modalities. The goal is to comprehensively train an end-to-end assistant for crop disease identification and prevention measures, simplifying its implementation in practical application environments.

- A multimodal large language model for the recognition of agricultural crop diseases and pests has been proposed. By employing the Low-Rank Adaptation (LoRA) fine-tuning method¹⁷, the existing ChatGLM model is extended. The model achieves precise recognition of crop pest and disease images and provides expert knowledge-based question answering.
- Through the Q-Former framework, the model transitions from a single modality of language to a multimodal image-language format, enhancing its ability to extract textual information and visual representation features, thereby maximally bridging the gap between modalities.
- Experiments demonstrate that the proposed LLMI-CDP model outperforms five state-of-the-art models in the domain of agricultural crop disease recognition and knowledge-based question answering. These models include VisualGLM¹⁸, QWen-VL¹⁹, VisCPM²⁰, MiniGPT4²¹, and Ziya-Visual²².

Related works

Language models

Modern pre-trained language models are predominantly built upon Transformer architectures like the GPT series and BERT, utilizing autoregressive Transformer models for scalable language model pre-training in a broad range of text corpora²³. Three forms of pre-training frameworks can be distinguished: encoder-decoder models, autoencoders, and autoregressive models. Through training on large-scale text corpora, LLMs have made significant progress and have become increasingly valuable in a variety of domains. The emergence of LLMs has triggered a paradigm shift in technology²⁴. Natural language processing (NLP) has made significant advancements, thanks to several open-source large-scale models, including LLaMA, BLOOM, and ChatGLM. Simultaneously, the multilingual language model ChatGLM-6B supports both English and Chinese. Utilizing technology similar to ChatGPT, ChatGLM-6B is optimized for Chinese question-answering and dialogue²⁵. It is comprehensively trained on an equal ratio of Chinese and English corpora, endowing ChatGLM with robust bilingual question-answering capabilities, supported by techniques like supervised fine-tuning and self-feedback. Despite having only 6.2 billion parameters, ChatGLM-6B can generate answers that align with human preferences²⁶.

Building upon the ChatGLM-6B model, this paper has enhanced the large language model's ability to capture image features through fine-tuning. The groundwork for developing multimodal language models is laid by this work. In this study, integration of domain-specific knowledge concerning agricultural crop diseases and pests into the ChatGLM-6B model occurs. Repositioning the base language model customizes it for a specialized corpus dedicated to the management of agricultural crop diseases and pests.

Vision-language models

Given the progress in LLMs and visual modeling, scholarly attention has been progressively directed towards visual LLMs. This heightened interest is reflected in the growing body of academic research dedicated to the utilization

of large prediction models for addressing multimodal tasks. Subsequently, the emergence of the concept of multimodal large language models (MLLMs) is a noteworthy development in the academic discourse²⁷. Various theories and methods have been developed to incorporate visual data into LLMs, fine-tuning these models with specific instructions to enhance their precision and performance. This approach has been demonstrated to enhance the adaptability of language models when addressing new tasks, significantly increasing their ability to generate textual information from visual inputs. In recent years, the focus of image and language research has shifted significantly, moving from a broad focus on language models to a more specific emphasis on models that integrate vision and language. GPT-4 V has demonstrated powerful performance across various tasks, capable of receiving multimodal inputs and providing detailed explanations based on customized instructions for different multimodal tasks²⁸. Although GPT-4 V has not been made open source yet, its robust capabilities have sparked a new wave of research, aiming to extend language models into the multimodal realm. This involves endowing LLMs with visual reasoning capabilities, similar to the approaches taken by models like MiniGPT-4, LLaVA, and LLaMA²⁹. MiniGPT4, through pre-training on 134 million image-text pairs, connects a frozen visual encoder with a LLM, and then enhances the model's performance by further fine-tuning on well-aligned image-text datasets³⁰. LLaVA also utilizes image-text pairs to calibrate the visual model and the large language model³¹. Unlike MiniGPT4, LLaVA fine-tunes the entire LLM on 150 K high-quality multimodal instructions generated by GPT-4. While these approaches demonstrate impressive multimodal understanding capabilities, they require updating billions of model parameters and meticulously collecting large amounts of multimodal training data. This data is either annotated by humans or extracted from responses of the OpenAI API.

Moreover, these models are primarily designed for general domains and have not been fine-tuned specifically for crop disease-related information, resulting in reduced precision of their generated responses. Our work aims to endow the foundational LLMs with the capability to understand visual attributes. In this context, our model introduces an innovative LoRA fine-tuning strategy, which includes fixing the inherent parameters of the initial pre-trained model. Enhancement is achieved by integrating auxiliary matrices to replicate the comprehensive fine-tuning of model parameters. This strategic implementation reduces computational demands and, propelled by low-rank adaptation, gradually integrates image-based visual attributes into the pre-existing ChatGLM model. Since existing precedent models have not yet attained the level of proficiency required for the agricultural crop disease domain, our model is set to exhibit a high level of competitiveness in the area of knowledge-based question answering for crop diseases, compared to previous multimodal models.

Language-image pre-training

Human understanding of the external environment primarily occurs through two fundamental channels: vision and language. The main challenge faced by models that integrate images and language is effectively combining these image features into a scalable language model capable of understanding image feature data. Currently, the adoption of the Transformer architecture has become the predominant approach in the field of multimodal algorithms³². This architecture effectively combines information from different modalities at a feature level comprehensible to LLMs, simplifying the process of feature fusion. BLIP-2 introduces a pre-training-based approach that enhances multimodal task performance through the joint training of visual and language models³³. By incorporating a Multimodal Encoder-Decoder structure, it effectively facilitates multitask pre-learning and transfer learning. In a range of vision-language tasks, including picture-text retrieval, image captioning, visual question answering, visual reasoning, and visual dialogue, BLIP-2 exhibits state-of-the-art performance³⁴. By leveraging pre-trained visual and language models, BLIP2 enhances multimodal effectiveness and reduces training costs. The pre-trained visual models provide high-quality visual representations, while the pre-trained language models offer robust language generation capabilities³⁵. To reduce computational costs and counteract catastrophic forgetting, there is an inclination to fix the parameters of the visual and language models in Vision-Language Pre-training. BLIP-2 is a versatile and efficient pre-training method that facilitates vision-language pre-training by utilizing frozen LLMs and readily available frozen pre-trained image encoders. As illustrated in Fig. 1, BLIP-2 bridges the modal gap with a lightweight Q-Former, which undergoes pre-training in two stages. The initial stage utilizes a frozen image encoder to guide the acquisition of visual-language representations. The second stage guides the learning of visual-to-language generation from a frozen language model. BLIP-2 achieves state-of-the-art performance on a range of vision-language tasks while using a significantly smaller number of trainable parameters compared to previous approaches. Due to the use of frozen unimodal models and the lightweight Q-Former, BLIP-2 is more computationally efficient than existing techniques, maximizing performance enhancement while minimizing computational costs.

The model established in this paper adopts the BLIP-2 pre-training strategy, which involves encoding and decoding images and text, followed by the fusion of their extracted features into the Q-Former framework. When it comes to extracting the most informative visual feature representations for textual content, Q-Former is an excellent choice. The combined data is then input into the large language model, where enhanced learning capabilities and dynamic adjustments ensure its adaptability. This strategy aims to improve consistency by simplifying the training process.

Low-rank adaptation

In the field of machine learning, the phenomenon of low-rank structures is widespread, with many machine learning algorithms inherently exhibiting low-rank characteristics³⁶. Moreover, in many deep learning tasks, especially those involving heavily over-parameterized neural networks, the trained neural networks often exhibit low-rank properties. Some early research even directly imposed low-rank constraints during the training process of the original neural networks. After fine-tuning language models for specific tasks, the weight matrices usually exhibit a very low intrinsic rank³⁷. Researchers believe that the amount of parameter updates, even when projected into a smaller subspace, does not compromise the effectiveness of learning. Thus, the approach

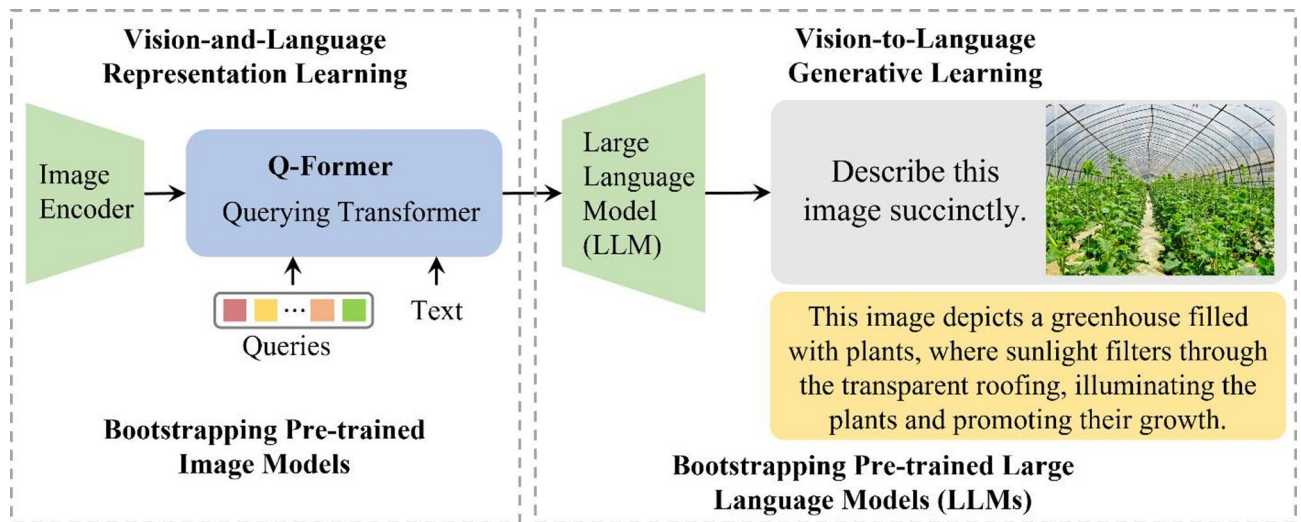


Fig. 1. Overview of BLIP-2's framework.

of fixing pre-trained model parameters and adding the product of low-rank matrices as trainable parameters alongside the original weight matrix has been proposed³⁸. This simulates the changes in parameters. The LoRA method can reduce the number of training parameters and graphics processing unit (GPU) memory usage while enabling the trained model to achieve performance comparable to full-scale fine-tuning. To significantly reduce the number of trainable parameters for downstream tasks, the weights of the pre-trained model are frozen, and trainable rank-decomposed matrices are inserted into each layer of the Transformer architecture.

In the fine-tuning of large language models, Prefix Tuning improves model generation by appending trainable prefix vectors to the input layer, but increasing the prefix length significantly increases GPU memory consumption. P-Tuning v2 enhances this by introducing trainable prompt parameters across multiple Transformer layers, improving task-specific representations at the cost of a 1–3% increase in total model parameters. However, both methods face critical limitations. Training stability and effectiveness are highly sensitive to initialization strategies and hyperparameters, with improper settings leading to convergence difficulties, especially in low-data scenarios. The choice of prefix length also presents a paradox: insufficient length fails to capture task semantics, while excessive length introduces redundancy and noise. Additionally, the shared prefix architecture constrains the model's expressiveness, particularly for complex tasks, and does not address the knowledge bias from pretraining.

While P-Tuning v2 improves upon its predecessor by employing hierarchical prompt injection, it faces challenges in multi-layer parameter co-optimization, leading to training instability and poor convergence, particularly under resource constraints. Both methods exhibit performance fluctuations in few-shot learning, with high sensitivity to data distribution shifts and limited interpretability. Despite being more memory-efficient than full-parameter fine-tuning, they still require intermediate state storage during gradient computation for long-sequence processing, offering limited hardware efficiency improvements. These observations highlight the trade-offs between parameter efficiency and semantic adaptation in prompt-based approaches, especially in mitigating knowledge bias for complex tasks.

In the model fine-tuning approach of this paper, the adaptive strategy of LoRA will be used to enhance the efficacy of fine-tuning LLMs for downstream tasks. It achieves this without increasing inference latency or shortening input sequences, while still maintaining excellent model performance. LoRA also demonstrates exceptional capability in service deployment scenarios, achieving rapid task switching by sharing most of the model parameters. Through global training approximation, this framework minimizes resource waste and maximizes performance.

To achieve optimal overall performance, LoRA ingeniously employs attention-related matrices, including W^q and W^v , while also considering W^k . The most significant advantage of LoRA is its faster speed and lower memory occupancy.

The LLMI-CDP model

The framework of the proposed LLMI-CDP model

Based on the VisualGLM model, the LoRA technique is utilized to fine-tune the model developed in this paper. The training process involves utilizing the created image-text data containing information on agricultural crop diseases and pests. During this procedure, pertinent parameters in the language model and image encoder stay fixed, while LoRA settings in both components are meticulously refined. Additionally, parameters related to the Q-Former are also adjusted. Consequently, this model demonstrates effective question-answering capabilities for the identification and management of specialized agricultural crop disease features. It also performs well in the extraction and recognition of features from pest images. The integration of a multimodal large language

model contributes to the advancement of research in the field of prevention and management of agricultural crop diseases and pests.

The LLMI-CDP model consists of five main components: image-text matching, image encoding, text encoding, LoRA fine-tuning, and result testing. These components collaborate to optimize model performance, as shown in Fig. 2. Image-text matching ensures a coherent relationship between images and their corresponding descriptions, which is essential for subsequent processing. Image encoding transforms raw image data into feature representations, while text encoding converts textual data, such as captions, into embeddings that capture semantic content. For efficient training, LoRA fine-tuning is used to freeze the image encoder's parameters and expand the language model's parameters. This method enhances the language model without retraining the entire system, conserving computational resources. The Q-Former connects the image encoder and the language model, facilitating effective interaction between the two modalities for accurate predictions. Result testing assesses model performance in tasks like image captioning and image-text retrieval, providing insights into its overall effectiveness. By combining these components, the LLMI-CDP model efficiently handles complex multimodal tasks while minimizing computational overhead.

Q-Former, as the core module for aligning visual and language modalities, aims to efficiently map the image features from the visual encoder to the semantic space of the language model through a lightweight cross-modal interaction mechanism. Its design consists of two stages: (1) Visual-Language Representation Learning Stage: Through cross-attention layers, Q-Former facilitates interaction between image features and learnable query vectors, generating visual feature representations that are semantically relevant to the text. This process enhances sensitivity to fine-grained visual attributes, such as lesion shape and color distribution. (2) Visual-to-Language Generation Stage: The optimized visual features are input into the language model, where its generative capabilities are employed to achieve image-text alignment. By dynamically adjusting the weights of the query vectors, Q-Former selects the most relevant visual features and suppresses redundant information. Compared to traditional methods that concatenate image and text features, Q-Former utilizes lightweight query vectors and attention weight training, avoiding large-scale adjustments to model parameters and reducing computational overhead. Additionally, it captures the relationship between local features and global semantics through bidirectional self-attention masking.

Given that other fine-tuning methods such as Prefix Tuning and P-tuning v2 did not yield satisfactory results in the fine-tuning training of this model, the study specifically utilizes LoRA for fine-tuning training. Therefore, the fine-tuned parameters include those related to LoRA in both the image encoder and the large language model, as well as relevant parameters in the Q-Former. The ultimate goal of this training process is to attain and retain a high level of multimodal proficiency in the LLMI-CDP model. Considering the consumption of hardware resources, this training method significantly reduces training costs and time. During the image encoding process, features of the images are learned for representation, ultimately encoding image attributes into feature vectors. During the text encoding phase, the primary objective is to utilize the generated vectors for comparative learning. The subsequent focus is on cross-attention fusion analysis with the resulting components.

Text vectors are generated by encoding the textual descriptions of images as part of the text encoding process. Thereafter, the vector dimensions are normalized through a residual layer to align them with the dimensions of the image vectors, facilitating comparative learning. The objective of image-text matching is to establish a detailed alignment between text and image representations. The model must determine whether a pair of images and text is positive (matched) or negative (unmatched) in this binary classification task. We employ bidirectional self-

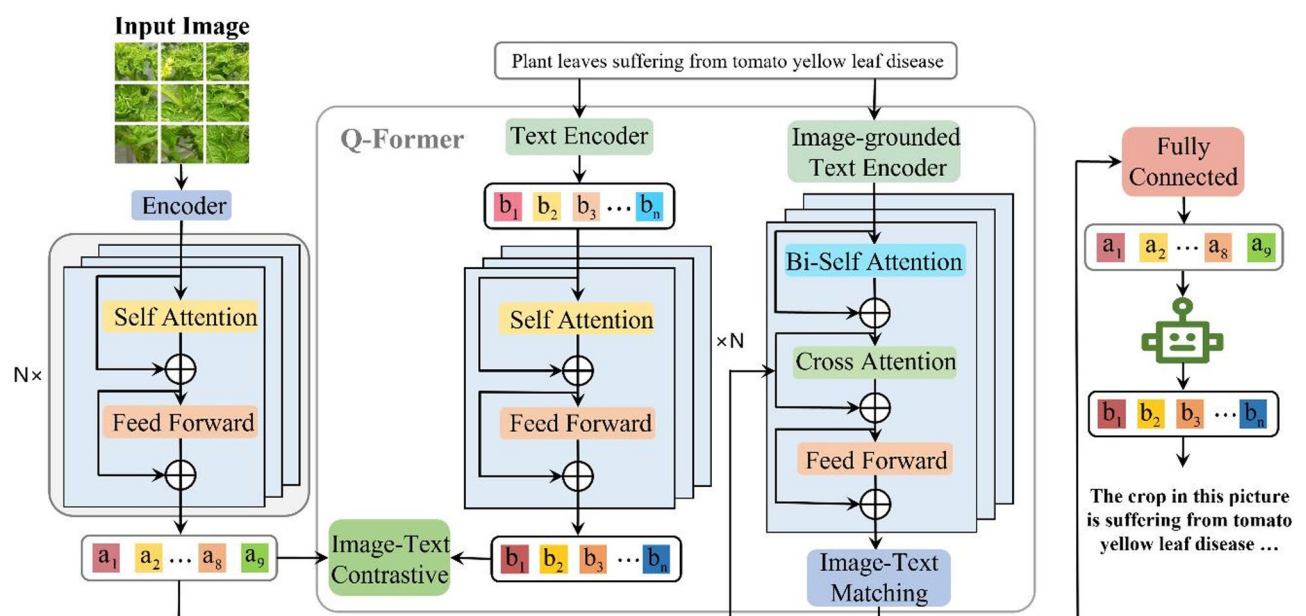


Fig. 2. The architecture of the LLMI-CDP model.

attention masks, allowing all queries and texts to attend to each other. As a result, the multimodal information is captured by the output query embeddings Z . Each output query embedding is fed into a binary linear classifier to obtain logits, and the logits of all queries are averaged to produce the output matching score.

In the result testing phase, a question and an image are input into our LLMI-CDP model. The new model then provides exemplary responses based on the acquired skills and the attributes of the input image. The following sections will detail the five components of the LLMI-CDP model.

LoRA fine-tuning training

The fine-tuning process using LoRA is divided into two phases, as shown in Fig. 3. In the first phase, the model focuses on training with input images. The images are encoded to extract features, which are then transformed into one-dimensional column vectors. These vectors are processed sequentially, from left to right and top to bottom, to form a structured representation of the image features. During this phase, the image encoder's frozen parameters are gradually fine-tuned, with LoRA optimizing the learning process. This approach enhances image feature quality without retraining the entire image encoder, saving computational resources. In the second phase, Q-Former enters the comparative learning stage, aligning image features with corresponding textual representations. The text vector matrix, containing descriptions related to each image, is concatenated with a base vector to create an enhanced matrix that matches the dimensionality of the text vectors. This step fine-tunes the text encoder's parameters, improving the integration between image and text encoders. By structuring the training process in these two phases, the model effectively learns to align and integrate image and text features, improving performance in tasks like image captioning and image-text retrieval.

In the Q-Former, textual features stems from text encoding, leading to the establishment of a matrix of textual feature proficiency. The skill matrix is aligned within the Q-Former, employing a cosine similarity-based strategy for the alignment process analysis. The aligned vector features are then inputted into the ChatGLM language model for training. Throughout the entire training process, the parameters of the ChatGLM language model remain fixed while being fine-tuned in conjunction with LoRA. The mathematical process of LoRA fine-tuning is described as follows.

$$W = W_{pm} + tW_{LoRA} = W_{pm} + tE_{LoRA-zeros} \times F_{LoRA-gaussian}, \quad (1)$$

In the above equation, t is a random variable whose absolute value does not exceed 1. W , W_{pm} , and W_{LoRA} represent the weight matrices of the training model, frozen model, and LoRA fine-tuning process, respectively. During model training, $F_{LoRA-gaussian}$ is initialized using a normal distribution, while the $E_{LoRA-zeros}$ matrix is initialized with zeros. In this way, when the training process begins, the frozen model will still be bypassed, resulting in a zero matrix.

In the process of optimizing LoRA, when applied to the Query and Value mapping matrices in the attention mechanism, an even greater fine-tuning effect is achieved. The weights of the Query and Value mapping matrices in the attention mechanism are determined using the following method.

$$W^Q = W_{pm}^Q + tW_{LoRA}^Q \quad (2)$$

$$W^V = W_{pm}^V + tW_{LoRA}^V \quad (3)$$

When fine-tuning LoRA and training image-text data Y through the multi-head self-attention layer, corresponding mappings generate the computation formulas for the Query matrix Q , Key matrix K , and Value matrix V . These formulas are as follows.

$$\text{Query: } Y \times W^Q = Y \times W_{pm}^Q + tY \times W_{LoRA}^Q \quad (4)$$

$$\text{Key: } Y \times W^K = Y \times W_{pm}^K + tY \times W_{LoRA}^K \quad (5)$$

$$\text{Value: } Y \times W^V = Y \times W_{pm}^V + tY \times W_{LoRA}^V \quad (6)$$

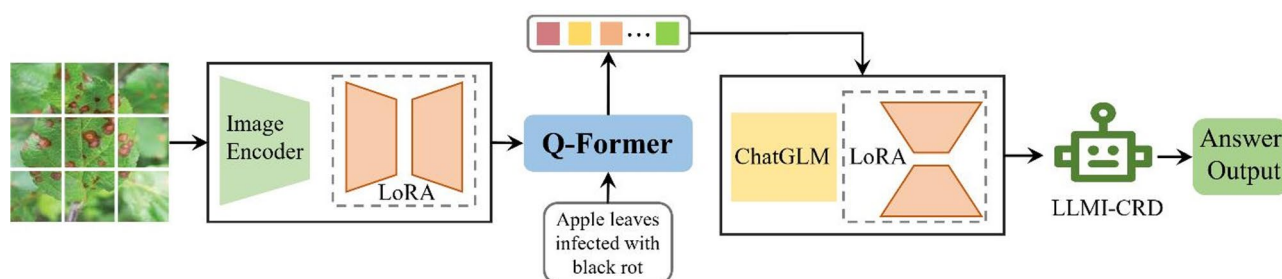


Fig. 3. The flow chart of the LoRA fine-tuning training.

When using the Softmax function, the computation inference for matrices Q and K with LoRA layers can be represented as follows.

$$\text{Softmax}(Q, K^T) = \text{softmax}\left(YW_{pm}^Q(W_{pm}^k)^T Y^T + tYW_{pm}^Q(W_{LoRA}^k)^T Y^T + tYW_{LoRA}^Q(W_{pm}^k)^T Y^T + t^2YW_{LoRA}^Q(W_{LoRA}^k)^T Y^T\right) \quad (7)$$

$$\text{Head} = \text{softmax}(Q, K^T) YW_{pm}^V + T \times \text{softmax}(Q, K^T) YW_{LoRA}^V \quad (8)$$

After undergoing LoRA fine-tuning, the final image-text comprehension skill matrix is formed and stored in the new model, thereby completing the LoRA fine-tuning process.

Image encoding and text encoding

In the image encoding process, the input image is first segmented into smaller blocks, each representing unique characteristics of the image. Every image is consistently divided into 14 by 14 area chunks. A comprehensive feature vector embedding is designed for each small block. After executing residual connection mapping through the self-attention layer, this process is further enhanced by a feed-forward network for residual connection mapping, creating feature column vectors. These feature vectors of the images are combined to form an enhanced feature matrix.

To prevent misaligned correlations, the vectors in this feature matrix are aligned with the text encoding vectors. A cross-attention mechanism is used to fuse and analyze the feature vectors of the images with the text. This method aims to further verify whether the image and text convey similar information, reducing alignment errors throughout the process. Additionally, it enhances the representation of feature vectors, enabling the model's text generation to provide an augmented representation.

During the text encoding process, the subject content in the image descriptions is assimilated and transformed into vector representations that reflect the dimensions of the image encoding. The text generates vector representations, denoted as $h_i (i = 1, 2, 3, \dots, 196)$. The complete description of the j -th image is aggregated into a feature matrix of text descriptions, represented as $H^j = [h_1, h_2, h_3, \dots, h_i, \dots, h_{196}]$. In the end, this feature matrix is combined with the subject vector to produce an enhanced text matrix, $H_a^j = [h_0, h_1, h_2, \dots, h_{196}]$.

After encoding, residual connections are established through the self-attention mechanism within the encoding part. These residual connections are created via a feed-forward network, involving vectors prior to the feed-forward network layer. This process generates a new enhanced text feature matrix, represented as $H_a^{j'} = [h'_0, h'_1, h'_2, \dots, h'_j, \dots, h'_{196}]$. It facilitates auxiliary analysis in comparative learning.

Image-text matching and answer testing

The role of image-text matching is to complete the fine-grained alignment learning between image and text representations. This enables the model to perform a binary classification task, predicting whether an image-text pair is matched or not. By using bidirectional self-attention masks, all queries and texts can mutually attend to each other. Since the image-text matching score can be fine-tuned on domain-specific datasets, it can make complex judgments about the multimodal interactions of the input using learned features.

After encoding the textual information, a text feature vector matrix is generated. Meanwhile, the subject content of the text is processed separately, yielding a specific vector. Finally, the text feature vector matrix is combined with this specific vector of subject content, creating a new enhanced feature matrix. Further cross-fusion and alignment of this matrix with the previously inputted picture feature vectors is performed. This ensures a finer granularity of alignment in the image-text matching task, guaranteeing maximum relevance of positive and negative samples during the matching process.

During the answer testing phase, images are encoded through an image encoder, generating image feature vectors. These feature vectors produced in the answer testing stage are denoted as j_k , where $k = 0, 1, 2, \dots, 196$. These features are then aggregated to construct the feature matrix of the image, represented as $J = [j_0, j_1, j_2, \dots, j_k, \dots, j_{196}]$. Simultaneously, queries intended to extract information from the image are input. These queries are encoded to form question vectors.

The Q-Former is then fed the question vectors and the picture feature matrix. The image-text information vectors are extracted to create a new matrix of vectors. Next, the feature matrix's dimensions are altered using a fully connected layer. Afterward, the modified matrix is input into the language model, which then generates responses to the presented questions. The language model generates responses by using feature data extracted from the input image throughout the entire response process, resulting in a textual representation.

Experiments

Dataset

To facilitate the acquisition of a high-quality multimodal agricultural dataset for model training, this study has established a multimodal agricultural crop disease and pest dataset in Chinese, comprising diverse images depicting various crop diseases and pests along with corresponding textual information. Each image of a pest or disease is labeled with the name of the corresponding condition. For each image, three to four or more relevant questions are generated to facilitate comprehensive learning through the integration of visual information. The textual answers primarily encompass information related to the diseases presented in the images, disease characteristics, and corresponding preventive and control methods. The dataset comprises a total of 2,498 color images depicting agricultural crop diseases, encompassing 141 categories of crop disease and pest types.

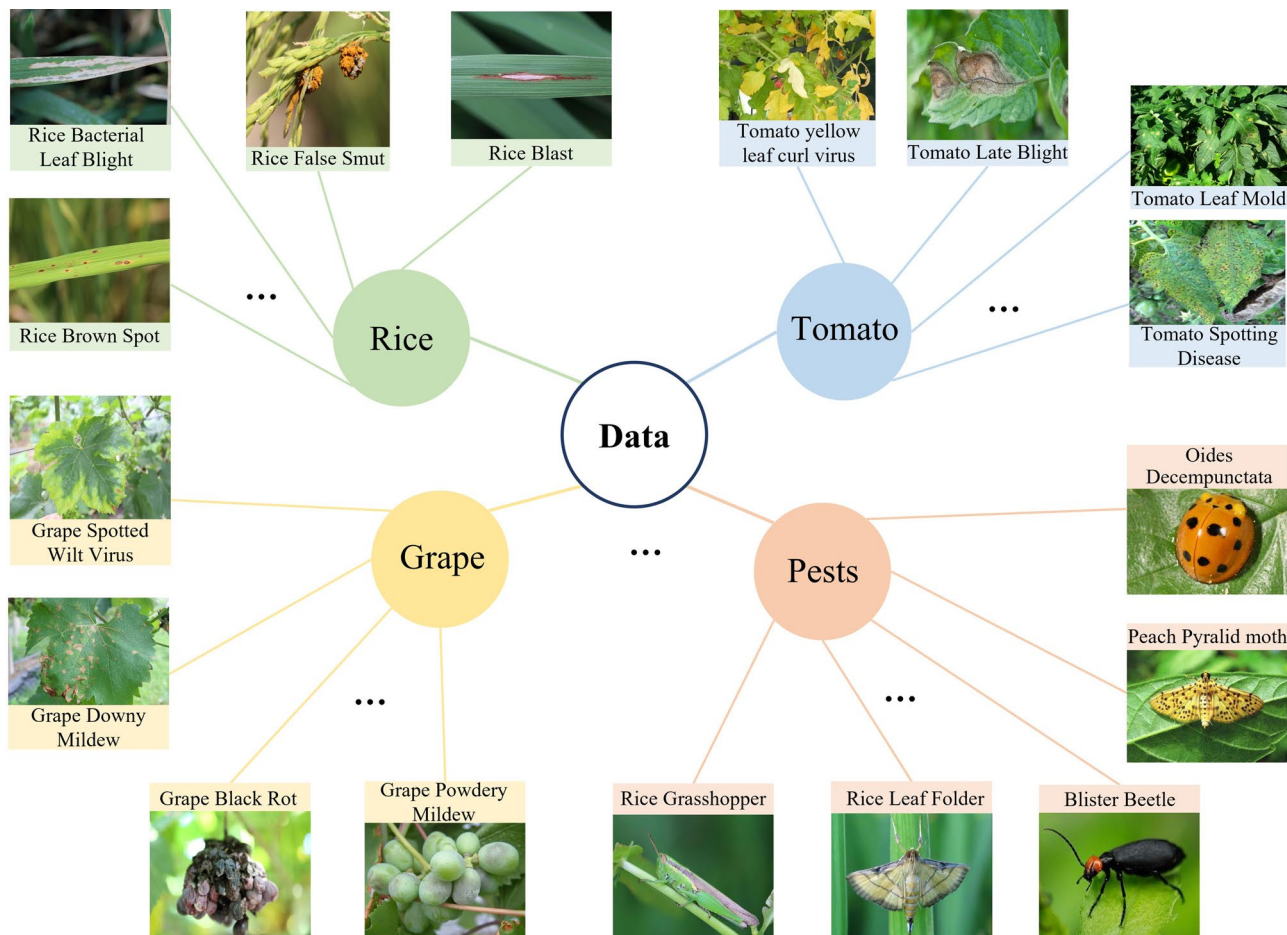


Fig. 4. Overview of some datasets.

Types	Crop category	Disease category	Total sample
Cereals	6	38	591
Vegetables	7	43	623
Fruit trees	4	28	462
Pests	-	32	822
Total	15	141	2498

Table 1. Dataset category distribution.

As depicted in Fig. 4, the collection of primary disease types for each crop includes information such as disease name, disease characteristics, and disease control measures. The gathered information on agricultural crop disease and pest images is categorized into four major classes: cereals, vegetables, fruit trees, and pests. These classes encompass a total of 15 crops and 32 pests, involving a comprehensive range of 141 disease categories. Detailed data for each category are presented in Table 1.

Comparison models

We evaluated five leading open-source MLLMs, including Ziya-BLIP2-14B-Visual, MiniGPT4, VisCPM, and Qwen-VL, as follows:

Based on the ChatGLM-6B language model, VisualGLM-6B utilizes a pretraining approach that leverages 30 million high-quality Chinese image-text pairs from the CogView dataset and 300 million carefully selected English image-text pairings. This training methodology effectively aligns visual information with the semantic space of ChatGLM.

VisCPM is a multimodal conversational model designed for bilingual dialogue with a focus on images in both Chinese and English. The model employs the Muffin visual encoding architecture and utilizes CPM-Bee (10B) as its language base model. The integration of visual and language models is achieved through language modeling training objectives. Leveraging the strong bilingual capabilities of the CPM-Bee base, VisCPM

demonstrates outstanding multimodal proficiency in Chinese through pretraining solely on English multimodal data, showcasing effective cross-lingual generalization.

Qwen-VL can generate text and detection boxes as outputs, and it can accept photos, text, and detection boxes as inputs. The language model in Qwen-VL is initialized with the pretrained Qwen-7B model. The Qwen-VL series is distinguished by its strong overall performance and sophisticated fine-grained recognition and understanding capabilities.

MiniGPT4 is a composite model constructed by combining the pretrained Vicuna Large Language Model (LLM) with the visual models VIT-G and Q-former. To enhance naturalness and usability, MiniGPT4 undergoes fine-tuning on a high-quality curated instruction dataset along with corresponding image and text pairs.

The Ziya-Visual multimodal large model is based on the training of the Ziya Universal Large Model V1 and possesses capabilities in visual question answering and dialogue. Drawing inspiration from excellent open-source implementations like Mini-GPT4 and LLaVA, Ziya-Visual enhances Ziya's image recognition capabilities. This integration enables Chinese users to experience the exceptional abilities of a large model that combines both visual and language modalities.

Experimental results

Numerous trials are conducted during the construction of the model. Every comparison test is conducted using an NVIDIA RTX 4090 GPU. The repetition penalty parameter is configured to 1.2, the temperature is set to 0.8, top-P is set to 0.4, and top-K is set to 100. The experimental tests mainly focus on crop disease identification tasks, as well as the description and management of crop disease characteristics. Images from both the constructed dataset and non-dataset sources are selected for experimental testing. During the LoRA fine-tuning of the model, we set the lora rank to 12. To ensure the effectiveness of the model in answering general questions post-fine-tuning, parameters in 4 random layers of the model were fine-tuned. The learning rate was set to 0.0001, the batch size to 4, and the number of training iterations to 10,000 steps. The changes in relevant indices during the fine-tuning training process are illustrated in the accompanying Fig. 5.

Cognitive task

For a multimodal large language model to achieve precise responses, the model needs to understand specific visual features based on instructions, align text with images, and utilize the knowledge of the large language model to generate responses. This presents a more intricate challenge than a singular image perception task. Applying MLLMs to resolve specialized problems requires in-depth exploration. The model needs to identify the fundamental questions and ensure the accuracy of the recognized content. Accurately identifying crop disease categories can effectively contribute to disease prevention and facilitate the management of crop diseases.

Perceptual recognition is a fundamental capability of MLLMs. In this experiment, a single-round dialogue format is used, where the model is presented with an image and asked: "What disease is affecting the crop in this image?". Figure 6a illustrates the results of various models, including the LLMI-CDP model. In this task, the two comparison models were unable to identify what type of crop leaves were in the picture. There was also a significant discrepancy in the identification of the disease type, failing to accurately recognize the disease affecting the crop in the image.

The Qwen-VL model could identify the leaf information of the crop in the image, but its answers regarding the disease type were not precise. The MiniGPT4 model demonstrated good effectiveness in extracting image features and had remarkable perceptual abilities for subtle features of crop leaves. Still, this model was not outstanding in identifying disease types. It showed limited capability in responding to queries in Chinese but performed relatively better with English questions. Through multiple validations and comparative experiments, the LLMI-CPD model demonstrates outstanding performance not only in the recognition of crop diseases but also in pest identification. In Fig. 6b, the performance of various models in pest identification tasks is presented. Models like VisualGLM are not very efficient in recognizing pest images. This inability to accurately identify pest categories is a significant drawback for the effective use of general visual large models in the agricultural sector. The LLMI-CPD model, developed in this paper, demonstrates proficiency in Chinese question-answering by providing precise and contextually relevant responses. Tables 2 and 3 show the recognition evaluation indicators of crop diseases and pests in the recognition task respectively. The model LLMI-CDP proposed in this article achieved the best results in answering questions about diseases and pests.

Question and answer

Agricultural crop disease management and prevention measures should be rational and effective. Providing scientific disease prevention and management strategies for crop growth has significant value for enhancing crop yield and ensuring quality. Accurate answers to related disease questions play a crucial role in large models. We organized questions related to crop disease management and prevention. The models were engaged in multiple rounds of dialogue to assess their answer capabilities, with experimental results shown in Fig. 7.

During the evaluation, answers provided by the models were assessed based on GPT-4. Nonetheless, relying solely on GPT-4 for assessing responses does not ensure the accuracy of the associated preventive and management measures. The influence of GPT-4's ambiguous replies to certain queries on the evaluation process cannot be overlooked. Consulting relevant professional literature and seeking advice from industry experts for manual evaluation of the models is also a good option. Therefore, the evaluation of the models involved both GPT-4 and human judgment. Setting the weight ratio of GPT-4 and human evaluation at 3:7, the evaluation scores, as depicted in Table 4, are obtained. Figure 8 visually depicts the performance evaluation scores of each model.

In this task, the Qwen-VL and VisCPM models demonstrated excellent performance in responding to instructions and proposing specific preventive measures, providing comprehensive answers. However, all the

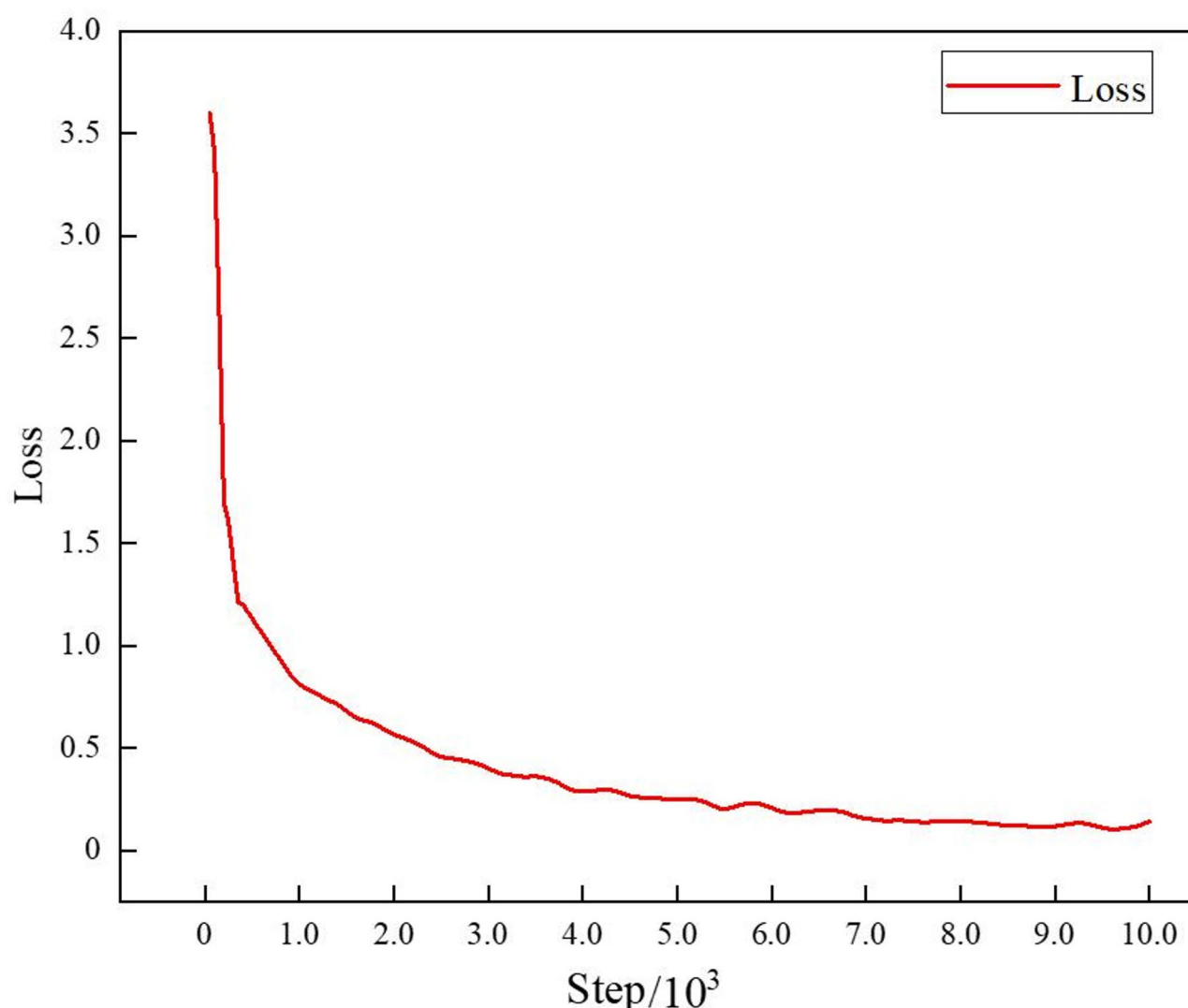


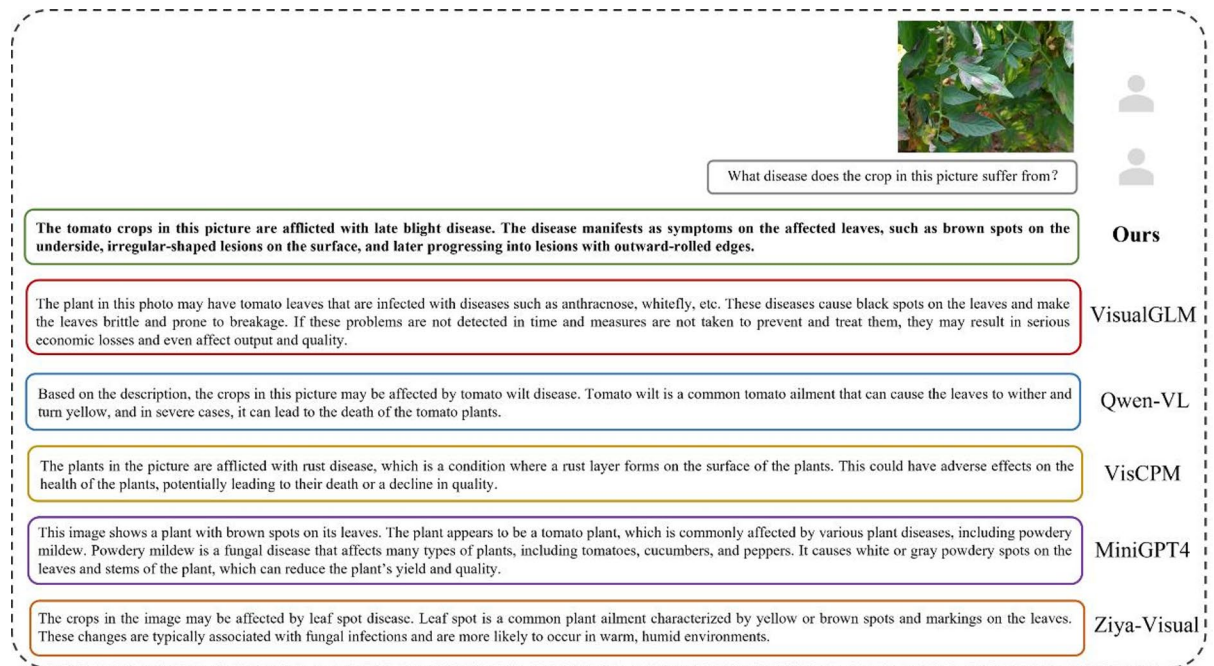
Fig. 5. Fine-tuning training loss variation curve.

models exhibited issues of lacking specificity in multi-round dialogue question-answering. The Ziya-Visual model could only offer limited prevention methods, and its answers were broadly similar for different diseases, lacking detailed explanations and guidance. The MiniGPT4 model was able to output more detailed prevention suggestions but lacked specific management measures.

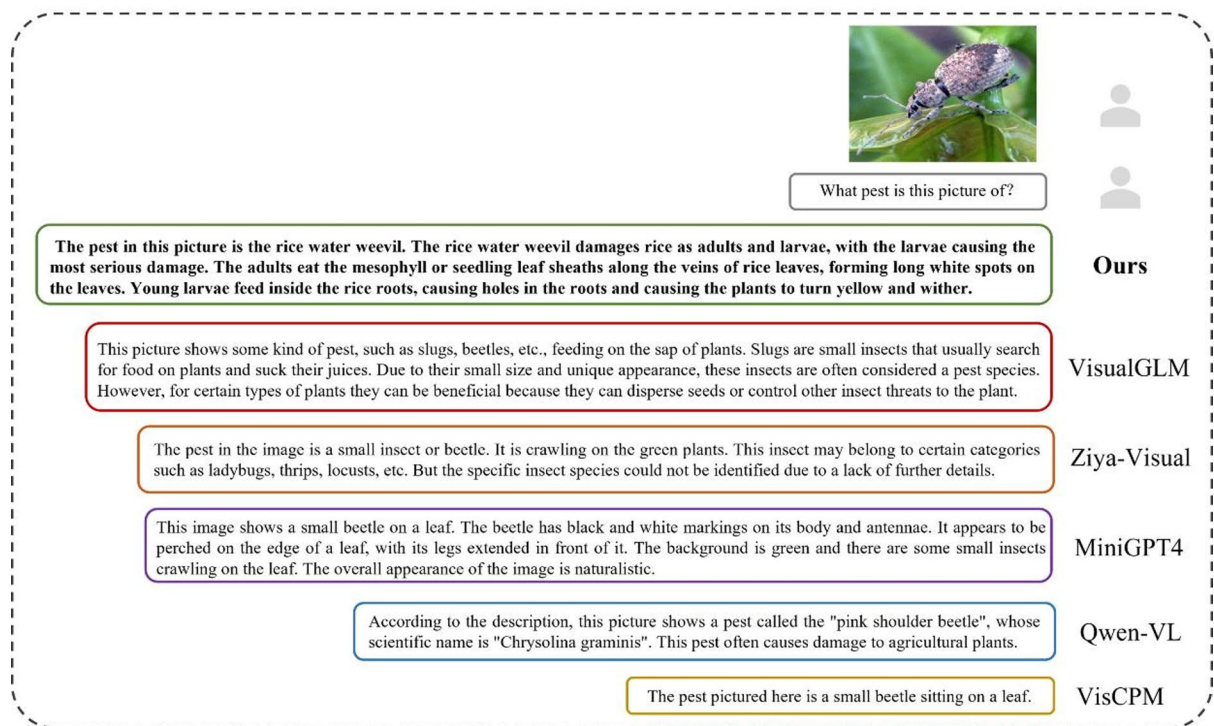
The model created in this paper excelled in providing more precise prevention methods, with proposed control measures being more comprehensive and accurate. Based on the assessment using GPT-4, our model's proposed solutions received excellent scores. For certain crop diseases, it was able to provide specific and targeted advice.

Conclusions

This paper presents a comprehensive multimodal language model (LLMI-CDP) designed for identifying crop pests and diseases. The model leverages a multimodal dataset focused on Chinese agricultural crop diseases and pests, incorporating both visual and textual information from crop images. Through a fine-tuning training process using LoRA, the model enhances its understanding and processing capabilities for both image and text data, making it highly effective for agricultural applications. Further optimization through additional fine-tuning training refines the performance of LLMI-CDP, improving its accuracy and robustness in recognizing and diagnosing crop diseases. LLMI-CDP represents a notable advancement in the development of practical agricultural assistants, particularly in the Chinese context, where it marks the extension of MLLMs into specialized vertical domains. Despite these advancements, the model still has considerable room for improvement, especially in deep reasoning capabilities, which are critical for providing more sophisticated diagnostic insights and decision-making support. In the experimental section, the model is compared to five open-source MLLMs, where it consistently outperformed other models in recognizing and managing crop diseases. LLMI-CDP excelled in accurately identifying specific crop diseases and providing detailed, contextually relevant preventive



(a) Comparison of agricultural crop disease recognition experiments



(b) Comparison of agricultural crop pest identification experiments

Fig. 6. Comparison of cognitive task experimental results.

recommendations. However, despite its strengths, the recognition of agricultural diseases still presents significant challenges, and the model's performance could be further enhanced by addressing these gaps.

In future research, we aim to broaden the diversity of crops included in the dataset by integrating more granular crop categories. We will also explore the possibility of training small-scale image detection and recognition models to automate the labeling of images within the same category. While the cascaded architecture that combines the visual encoder and text decoder is efficient in terms of parameter usage, its inference latency remains constrained by the two-stage sequential computation process. This limitation poses challenges in meeting the real-time

Model	Accuracy (%)	Recall (%)	F1-score (%)
Qwen-VL	61.2	61.4	61.3
VisCPM	52.8	50.6	51.7
MiniGPT4	36.7	36.6	36.6
Ziya-Visual	25.5	21.2	22.1
VisualGLM	72.4	70.6	71.5
LLMI-CDP	78.8	75.4	77.1

Table 2. Comparison of model disease recognition performance in cognitive tasks.

Model	Accuracy (%)	Recall (%)	F1-score (%)
Qwen-VL	48.6	47.2	47.9
VisCPM	32.9	33.6	33.2
MiniGPT4	56.6	56.1	56.3
Ziya-Visual	42.8	42.7	42.7
VisualGLM	67.4	66.5	66.9
LLMI-CDP	86.7	86.4	86.5

Table 3. Comparison of model pests recognition performance in cognitive tasks.

requirements of edge devices deployed in agricultural environments. Future work will focus on addressing these challenges through the exploration of dynamic multimodal alignment networks, domain-adaptive incremental learning techniques, and lightweight heterogeneous model distillation. These approaches aim to overcome the current limitations in model generalization, task compatibility, and computational efficiency. Additionally, we will prioritize advancements in deep reasoning and contextual understanding to enhance the model’s capacity to deliver more nuanced and actionable insights for agricultural practitioners.

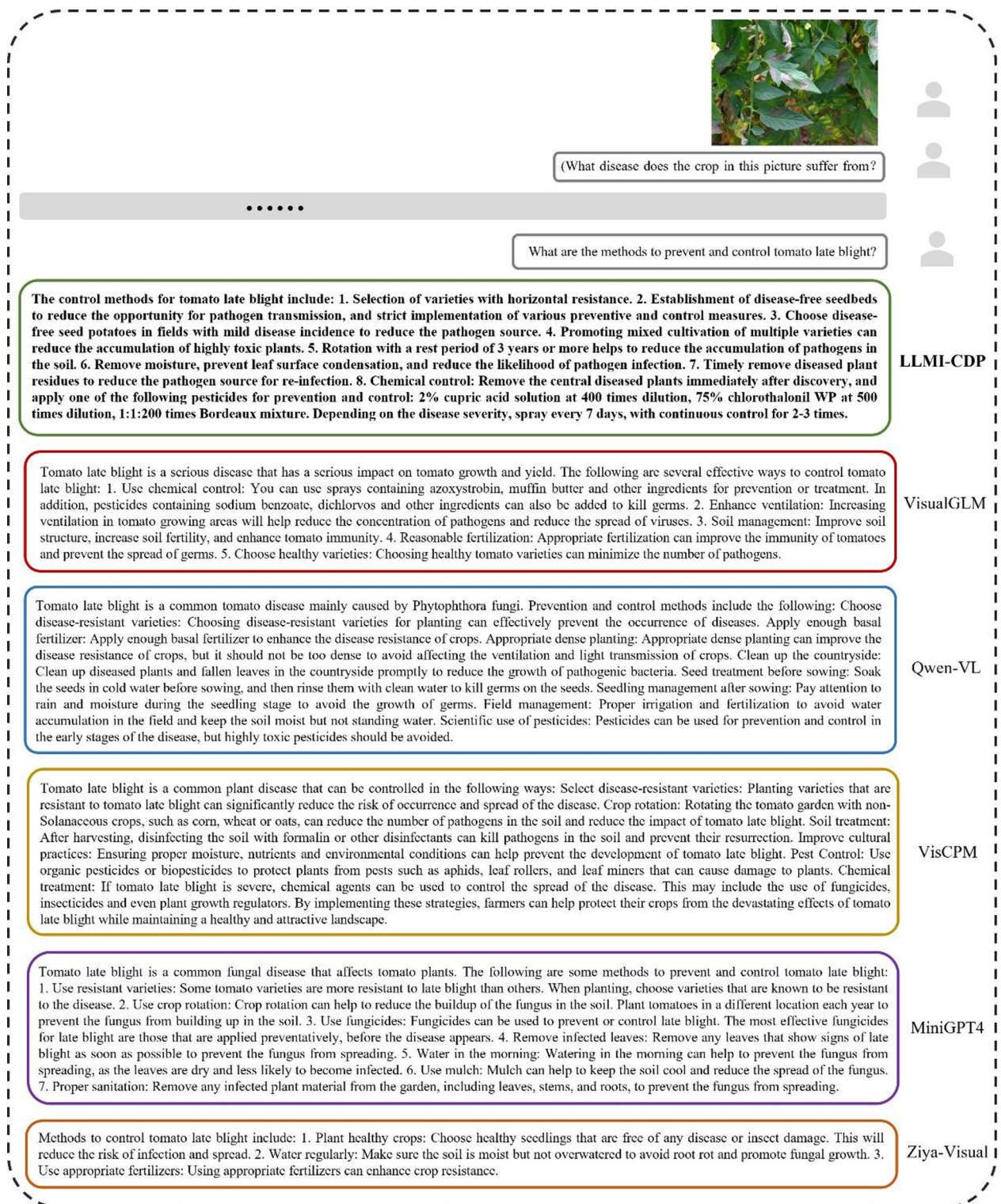


Fig. 7. Comparison of question-answering experimental results.

Model	Fluency (GPT-4)	Accuracy (GPT-4)	Fluency	Accuracy	Response speed	Total score
Qwen-VL	83.4	66.2	84	86	85	83.3
VisCPM	79.6	62.4	80	76	88	77.6
MiniGPT4	82.1	70.5	86	78	88	81.5
Ziya-Visual	62.8	45.6	67	65	82	65.2
VisualGLM	85.2	78.6	85	88	90	85.9
LLMI-CDP	89.5	88.7	88	90	90	89.1

Table 4. Performance evaluation details of each model.

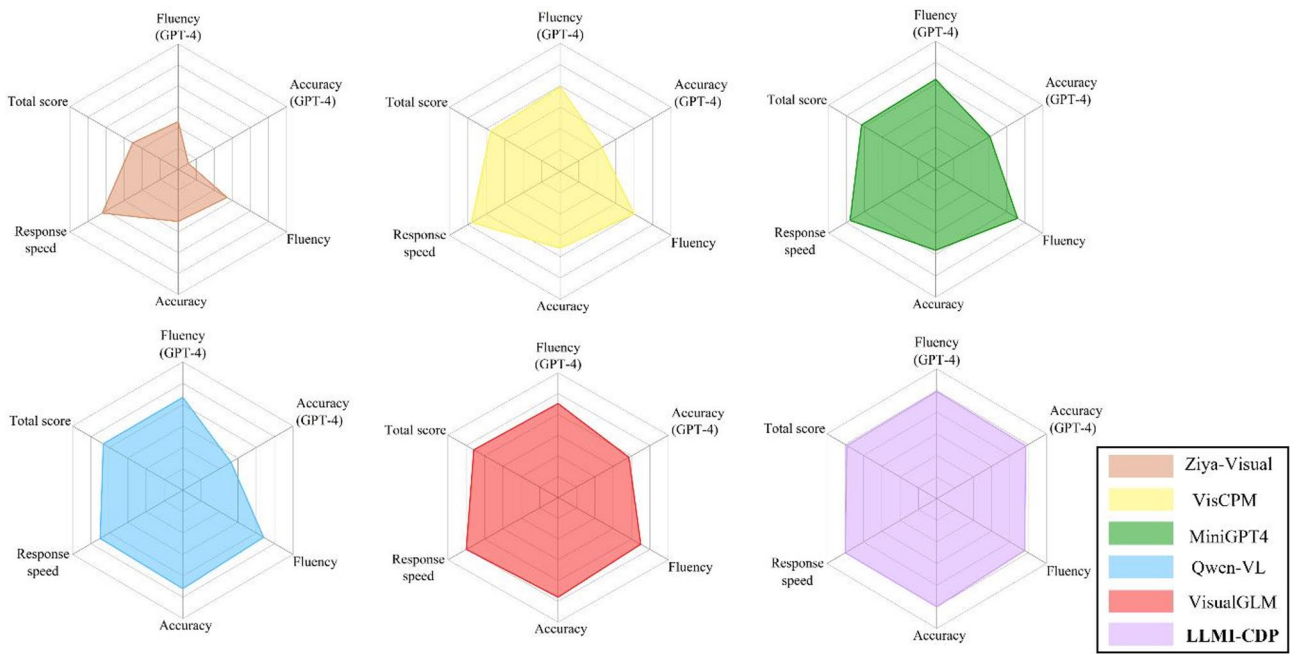


Fig. 8. Comparison chart of model evaluation scores.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 24 April 2024; Accepted: 9 May 2025

Published online: 01 July 2025

References

1. Yang, X. et al. A survey on smart agriculture: Development modes, technologies, and security and privacy challenges. *IEEE/CAA J. Automat. Sin.* **8** (2), 273–302 (2021).
2. Min, B. et al. Recent advances in natural Language processing via large pre-trained Language models: A survey. *ACM Comput. Surv.* **56** (2), 1–40 (2023).
3. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2018). Improving language understanding by generative pre-training.
4. Touvron, H. et al. Llama: Open and efficient foundation language models. arXiv:2302.13971. (2023).
5. Chen, Z., Balan, M., Brown, K. & M., & *Language Models Are Few-shot Learners for Prognostic Prediction* arXiv-2302 (arXiv e-prints, 2023).
6. Du, Z. et al. Glm: General language model pretraining with autoregressive blank infilling. arXiv:2103.10360. (2021).
7. Chowdhery, A. et al. Palm: Scaling Language modeling with pathways. *J. Mach. Learn. Res.* **24** (240), 1–113 (2023).
8. Cui, J., Li, Z., Yan, Y., Chen, B. & Yuan, L. Chatlaw: Open-source legal large Language model with integrated external knowledge bases. arXiv:2306.16092. (2023).
9. Xiong, H. et al. Doctorglm: Fine-tuning your Chinese Doctor is not a herculean task. (2023). arXiv:2304.01097.
10. Wang, H. et al. Knowledge-tuning large language models with structured medical knowledge bases for reliable response generation in Chinese. arXiv:2309.04175. (2023).
11. Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. arXiv:230709288. (2023).
12. Gulzar, Y., Ünal, Z., Kizildeniz, T. & Umar, U. M. Deep learning-based classification of alfalfa varieties: A comparative study using a custom leaf image dataset. *MethodsX* **13**, 103051 (2024).
13. Alkanan, M., Gulzar & Y. Enhanced corn seed disease classification: Leveraging MobileNetV2 with feature augmentation and transfer learning. *Front. Appl. Math. Stat.* **9**, 1320177 (2024).

14. Amri, E. et al. Advancing automatic plant classification system in Saudi Arabia: Introducing a novel dataset and ensemble deep learning approach. *Model. Earth Syst. Environ.* **10** (2), 2693–2709 (2024).
15. Gan, Z. et al. Vision-language pre-training: Basics, recent advances, and future trends. *Found. Trends® Comput. Graph. Vis.* **14** (3–4), 163–352. (2022).
16. Li, J. et al. Align before fuse: Vision and Language representation learning with momentum distillation. *Adv. Neural. Inf. Process. Syst.* **34**, 9694–9705 (2021).
17. Hu, E. J. et al. Lora: Low-rank adaptation of large language models. ArXiv Preprint arXiv:210609685. (2021).
18. Wang, W. et al. Cogvlm: Visual expert for pretrained language models. arXiv:2311.03079. (2023).
19. Bai, J. et al. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv :230812966. (2023).
20. Hu, J. et al. Large multilingual models Pivot zero-shot multimodal learning across languages. arXiv:230812038. (2023).
21. Zhu, D., Chen, J., Shen, X., Li, X. & Elhoseiny, M. Minigpt-4: Enhancing vision-language Understanding with advanced large Language models. arXiv:230410592. (2023).
22. Lu, J. et al. Ziya-VL: Bilingual large vision-language model via multi-task instruction tuning. arXiv:2310.08166. (2023).
23. Vaswani, A. et al. Attention is all you need. *Adv. Neural. Inf. Process. Syst.* **30**. (2017).
24. Liu, X. et al. Large language models are few-shot health learners. arXiv:2305.15525. (2023).
25. Zhang, Z., Tang, H. & Xu, Z. Fatigue database of complex metallic alloys. *Sci. Data.* **10** (1), 447 (2023).
26. Liang, P. et al. Holistic evaluation of Language models. arXiv :221109110. (2022).
27. Driess, D. et al. Palm-e: An embodied multimodal Language model. arXiv :230303378. (2023).
28. Zhang, R. et al. Llama-adaptor: efficient fine-tuning of Language models with zero-init attention. (2023). arXiv:2303.16199.
29. Lu, J., Goswami, V., Rohrbach, M., Parikh, D. & Lee, S. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10437–10446). (2020).
30. Gong, T. et al. Multimodal-gpt: A vision and Language model for dialogue with humans. (2023). arXiv:2305.04790.
31. Askell, A. et al. A general language assistant as a laboratory for alignment. arXiv :211200861. (2021).
32. Huang, H. et al. ChatGPT for shaping the future of dentistry: The potential of multi-modal large Language model. *Int. J. Oral Sci.* **15** (1), 29 (2023).
33. He, W. et al. Using augmented small multimodal models to guide large Language models for multimodal relation extraction. *Appl. Sci.* **13** (22), 12208 (2023).
34. Li, J., Li, D., Savarese, S. & Hoi, S. Blip-2: Bootstrapping Language-image pre-training with frozen image encoders and large Language models. arXiv :230112597. (2023).
35. Zhang, H., Li, X. & Bing, L. Video-llama: An instruction-tuned audio-visual Language model for video Understanding. arXiv :230602858. (2023).
36. Fang, Y. et al. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 19358–19369). (2023).
37. Ding, N. et al. Parameter-efficient fine-tuning of large-scale pre-trained Language models. *Nat. Mach. Intell.* **5** (3), 220–235 (2023).
38. Liu, X. et al. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. (2021). arXiv:2110.07602.

Acknowledgements

This work was supported in part by the Major Project of Scientific and Technological Innovation 2030 (2021ZD0113603), the National Natural Science Foundation of China (62276028), the Major Research Plan of the National Natural Science Foundation of China (92267110), and the Qin Xin Talents Cultivation Program, Beijing Information Science & Technology University (QXTCP A202102).

Author contributions

Y.W.: Conceptualization; Writing- Reviewing and Editing. F. W.: Writing - Original Draft; Conceptualization and Methodology; Visualization. W.C.: Conceptualization; Supervision and Funding acquisition. B.L.: Writing- Reviewing and Editing. M. L.: Writing- Reviewing and Editing. X.K. and Z.P.: investigation. C.Z.: Conceptualization; Writing- Reviewing and Editing. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025