Visual-Language Transformer-Based Tomato Leaf Disease Detection for Portable Greenhouse Monitoring Device

Manveen Kaur
University of Windsor
Rajmeet Singh
rajmeet.bhourji@ku.ac.ae
Khalifa University of Science and Technology
Shahpour Alirezaee
University of Windsor
Irfan Hussain
Khalifa University of Science and Technology

Research Article

Keywords: Tomato leave, disease, BLIP-2, LoRA, VLM, LLM

Posted Date: July 21st, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7077052/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Plant Methods on October 28th, 2025. See the published version at https://doi.org/10.1186/s13007-025-01456-8.

Visual-Language Transformer-Based Tomato Leaf Disease Detection for Portable Greenhouse Monitoring Device

Manveen Kaur[1†], Rajmeet Singh[2*†], Shahpour Alirezaee[1], Irfan Hussain[2*]

[1]Department of Mechanical Engineering, University of Windsor, Windsor, Canada.

[2]Department of Mechanical Engineering, Khalifa University, Abu Dhabi, UAE.

*Corresponding author(s). E-mail(s): rajmeet.bhourji@ku.ac.ae; irfan.hussain@ku.ac.ae;
Contributing authors: kaur79@uwindsor.ca; s.alirezaee@uwindsor.ca;
†These authors contributed equally to this work.

Abstract

Tomato leaf diseases pose a significant threat to global food security, necessitating accurate and efficient detection methods. This paper introduces the Tomato Leaf Disease Visual Language Model (TLDVLM), a novel approach based on the BLIP-2 architecture enhanced with Low-Rank Adaptation (LoRA), for precise classification of 10 distinct tomato leaf diseases. Our methodology integrates a sophisticated image preprocessing pipeline, utilizing GroundingDINO for robust leaf detection and SAM-2 for pixel-level segmentation, ensuring that the model focuses solely on relevant plant tissue. The TLDVLM leverages the powerful multimodal understanding of BLIP-2, with LoRA applied to its Q-Former module, enabling parameter-efficient fine-tuning without compromising performance. Comparative experiments demonstrate that the TLDVLM significantly outperforms baseline models, including CLIP-LoRA and ConvNeXT-tiny, achieving an accuracy of 97.27%, a precision of 0.9587, a recall of 0.9789, and an F1-score of 0.9681. Beyond classification, the finetuned TLDVLM checkpoints are integrated into a practical application for new image inference. This application displays the raw and segmented images, the predicted disease, and offers functionalities to fetch comprehensive information on disease causes and remedies using external APIs (e.g., OpenAI), with an option to download a

PDF summary for offline access on a portable device. This research highlights the potential of LoRA-adapted Vision-Language Models in developing highly accurate, efficient, and user-friendly agricultural diagnostic tools.

Keywords: Tomato leave, disease, BLIP-2, LoRA, VLM, LLM

Introduction

Crop diseases significantly threaten global food security, impacting staple crops like wheat and tomatoes, which are vital to agricultural economies. In Canada, greenhouse cultivation contributes approximately 294,000 tonnes of tomatoes annually, occupying over 6,200 ha and generating more than 600 million CAD revenue [8]. On contrary, Tomato production in the United Arab Emirates (UAE) has advanced significantly through the adoption of controlled-environment agriculture (CEA) and high-tech greenhouse systems designed to address the country's arid climate, high temperatures, and limited freshwater availability. Commercial operations such as Pure Harvest Smart Farms have demonstrated year-round production of up to 30 tomato varieties within 16-hectare climate-controlled greenhouses, achieving yields exceeding 400t/ha [1].

Tomatoes *(Lycopersicon esculentum)*, cultivated across 4.582 million hectares globally, face similar challenges from diseases such as early and late blight, leading to substantial yield reductions [33]. Traditional disease detection methods, including manual inspection by farmers and laboratory-based diagnostics, are labor-intensive, time-consuming, and often impractical for large-scale farming operations [22], [17], and [24].

Over the past decade, conventional machine vision techniques have been utilized for crop disease identification, addressing limitations of manual detection methods while reducing labor expenses. These systems primarily depend on computational algorithms to segment, extract features from, and classify images of diseased crop areas [13], [32]. A K-means clustering algorithm integrated with Particle Swarm Optimization (PSO) was introduced in [3] for segmenting early blight lesions in tomato images that were transformed into hue-saturation-value (HSV) color space, achieving an F1 score of 0.90. Nevertheless, since conventional machine vision technology employs a multi-stage algorithmic approach requiring human input, the performance of this methodology is heavily dependent on human expertise and design decisions.

As deep learning technology has advanced rapidly, convolutional neural networks (CNNs) have gained widespread adoption in agricultural disease detection [20]. The YOLOv3 (You Only Look Once) model was developed to identify 12 different diseases using 15,000 tomato images captured under natural field conditions, ultimately achieving a detection accuracy of 92.39%.A U-Net based segmentation approach was introduced in [25] for analyzing 1408 tomato leaf images, demonstrating an F1 score of 0.91. However, these models often struggle with real-world challenges, including variations in lighting, complex backgrounds, and the need to localize specific regions of interest (e.g., diseased leaves) within an image. To address these limitations, object detection and segmentation models, such as GroundingDINO and Segment Anything

Model 2 (SAM-2), have been developed to provide precise localization and segmentation of objects, enhancing the ability to isolate relevant features for downstream tasks [21], [27]. GroundingDINO leverages text prompts to detect objects with high accuracy, while SAM-2 generates detailed segmentation masks, enabling focused analysis of specific regions [12]. These advancements are particularly relevant for tomato disease detection, where identifying and isolating affected leaves is critical for accurate diagnosis.

Vision-Language Models (VLMs), which combine visual and textual processing capabilities, have emerged as a powerful tool for tasks requiring nuanced understanding of images and contextual information, offering a promising approach for agricultural applications [18]. Vision-Language Models, such as BLIP-2, integrate visual and textual data to perform tasks like image captioning and visual question answering, making them well-suited for interpreting complex agricultural imagery in conjunction with disease-specific prompts [5]. By fine-tuning BLIP-2 on a dataset of tomato leaf images annotated with disease labels, the model can learn to classify diseases based on visual patterns and contextual queries, such as "What disease is present in this tomato leaf?" [36]. This approach not only improves classification accuracy but also allows for interpretable outputs that align with human expertise, facilitating adoption in agricultural settings. However, deploying VLMs in real-world scenarios presents challenges, including the need for robust model fine-tuning, handling large model checkpoints, and managing computational resources efficiently [23]. Cloud-based platforms, such as Google Colab for computation, address these challenges by providing scalable infrastructure for storing and processing large models like BLIP-2.

This research introduces an innovative pipeline for tomato disease detection using a visual language model (TLDVLM) that combines GroundingDINO for tomato leaf detection, SAM-2 for segmentation, and a fine-tuned BLIP-2 using LoRA model for disease classification. The workflow is executed on Google Colab, where it retrieves the fine-tuned BLIP-2 model and processes input images to segment and detect tomato leaves before classifying diseases through a text-guided methodology. The model output provides both concise and detailed disease descriptions along with suitable treatment recommendations. The proposed TLDVLM system is implemented on a portable device designed for greenhouse environment applications. This research contributes to the field by demonstrating a scalable, automated solution for tomato disease detection, leveraging the strengths of VLMs to improve accuracy and interpretability, with potential applications in precision agriculture. The contributions of this work include:

- **Multi-Modal Vision-Language Architecture:** We developed a novel end-to-end pipeline integrating GroundingDINO object detection, SAM-2 instance segmentation, and fine-tuned BLIP-2 visual-language model, establishing a comprehensive framework that bridges computer vision and natural language processing for agricultural disease diagnostics.

- **Text-Guided Classification with Semantic Reasoning:** We implemented a text-conditioned disease classification system that leverages the semantic understanding capabilities of vision-language models to perform contextual disease identification, moving beyond traditional feature-based classification to interpretable, reasoning-driven predictions.
- **Edge-Deployable VLM Framework:** We design and deployed a computationally efficient visual language model on portable hardware for real-time inference in greenhouse environments, demonstrating the feasibility of deploying large-scale VLMs in resource-constrained agricultural settings.
- **Automated Disease Interpretation Pipeline:** We created an end-to-end automated system that generates structured disease descriptions and treatment recommendations without human intervention, addressing the semantic gap between visual disease detection and actionable agricultural insights through natural language generation.

Related Work

This section provides an extensive examination of existing literature on tomato disease classification. It further explores the implementation of vision-language models within agricultural applications and identifies current research limitations and opportunities for advancement.

Existing methods for tomato disease classification

Traditionally, tomato disease diagnosis has depended on manual assessment by agricultural specialists, an approach that is labor-intensive and prone to subjective interpretation errors. Contemporary developments in machine learning and deep learning methodologies have substantially enhanced the accuracy and operational efficiency of disease identification systems in tomato production.

Conventional machine learning approaches have been widely employed for tomato disease diagnosis, typically incorporating manual feature extraction procedures coupled with classification through diverse algorithmic methods. Noteworthy approaches include Zhang et al. [35], who employed color and texture features combined with support vector machines for tomato leaf disease classification, achieving reasonable accuracy on controlled datasets. Brahimi et al. [7] developed a system using traditional feature extraction methods including local binary patterns and color histograms, followed by random forest classification for identifying tomato diseases. Kaur et al. [14] leveraged traditional computer vision techniques combined with artificial neural networks to detect early and late blight in tomato leaves, demonstrating improved accuracy compared to manual inspection methods. Although these machine learning-based methodologies have demonstrated effectiveness, they frequently encounter limitations associated with the tedious and resource-intensive nature of manual feature extraction, leading to a transition toward deep learning-based solutions.

Deep learning methodologies have gained considerable attention in tomato disease diagnosis through the automation of feature extraction processes, delivering enhanced accuracy and improved system robustness. A deep learning model using faster R-CNN and single shot multibox detector (SSD) to enhance the accuracy of tomato disease detection under real-world acquisition conditions was proposed in [9], and achieving significant improvements over traditional approaches. Tm et al. [31] evaluated several CNN-based models including AlexNet and VGG16 with different training strategies to identify common tomato diseases including early blight, late blight, and bacterial spot. A system with deep learning classifiers, particularly focusing on transfer learning with pre-trained models, to accurately detect and classify tomato diseases has been proposed in [6], achieving high accuracy and demonstrating potential for mobile deployment. Abbas et al. [4] developed a DenseNet-based methodology for tomato plant disease classification, presenting a computationally efficient and precise model capable of real-time disease detection that surpassed existing approaches across various tomato disease datasets while demonstrating strong deployment feasibility. In [2], researchers proposed an integrated framework combining deep learning with ensemble learning techniques for tomato plant disease classification, achieving superior accuracy through advanced image preprocessing and the collaborative utilization of multiple CNN architectures. These investigations reveal both the advantages and constraints of deep learning models, emphasizing the necessity for more resilient and adaptable solutions that can effectively manage real-world variations encountered in tomato cultivation settings.

Recent advances have focused on more sophisticated architectures and training strategies. An AI-powered mobile application for tomato disease detection using deep learning models optimized for edge deployment, addressing practical deployment challenges in agricultural settings was proposed in [29]. Kawasaki et al. [15] improved tomato disease detection in greenhouse environments using an enhanced EfficientNet model, achieving high classification accuracies by analyzing environmental factors and optimizing model performance for controlled agricultural environments. These recent developments showcase the evolution towards more sophisticated, practically deployable solutions for tomato disease detection.

Vision-language models in agriculture

Vision-language models (VLMs) represent the forefront of artificial intelligence advancement, revolutionizing diverse applications through their ability to process and generate human-like text while simultaneously comprehending visual information. Built upon powerful architectures like CLIP [28] and BLIP [19], these models utilize massive datasets to learn joint representations of visual and textual information.

In the field of agriculture, VLMs' utilization has begun to be explored for various applications. PLLaMa, an open-source large language model based on LLaMA-2, specifically designed for plant science-related topics, demonstrating the potential of domain-specific language models in agricultural contexts was proposed in [34]. Kuska et al. [16] created AI-chatbots for agriculture to aid in interpreting decision support models in plant disease management, showcasing the integration of conversational AI with agricultural expertise. An application of large language models for answering agriculture-related questions, demonstrating significant promise in knowledge dissemination and decision support was proposed in [30]. Agriculture, with its unique environmental variables, disease manifestations, and regional nuances, requires datasets specifically tailored to local conditions and crop-specific characteristics for optimal accuracy. Existing works utilizing models like CLIP and BLIP in agricultural contexts highlight the need for datasets that accurately represent the diversity of real-world complexities inherent in tomato cultivation and disease manifestation.

A significant gap exists in incorporating VLMs into the operational workflows of agricultural practitioners, particularly those engaged in tomato cultivation. The agricultural workforce, characterized by varying technological expertise and widespread geographical distribution, requires intuitive, user-friendly interfaces and tools accessible under field conditions. Minimal research has addressed how VLMs can seamlessly integrate with agricultural workers to provide immediate, on-site diagnostic capabilities or management recommendations tailored specifically for tomato diseases.

Current VLM applications in agriculture often focus on general plant identification or broad agricultural question-answering, yet tomato disease diagnosis in field conditions demands a specialized approach that integrates domain-specific visual understanding with expert knowledge. The integration of vision foundation models like SAM (Segment Anything Model) with VLMs, as seen in precision farming applications, remains sparsely explored for tomato disease diagnosis. The potential of multimodal systems to combine detailed visual analysis of leaf symptoms with textual descriptions of disease progression and environmental factors for precise diagnoses represents an unexplored frontier. Furthermore, the application of parameter-efficient fine-tuning techniques such as Low-Rank Adaptation (LoRA) to VLMs for agricultural applications has received limited attention. These techniques offer the potential to adapt large-scale vision-language models to specific agricultural domains while maintaining computational efficiency, making them suitable for deployment in resource-constrained agricultural environments. The segmentation and localization of diseased areas within tomato leaves represents another underexplored area in VLM applications. While object detection approaches have been investigated, the integration of advanced segmentation models with VLMs for precise disease localization and classification has not been thoroughly explored in the context of tomato disease detection.

In summary, the literature reveals a distinct deficiency in utilizing VLMs for tomato disease diagnosis, further complicated by the requirement for agriculture-specific datasets, intuitive and accessible user interfaces, computationally efficient adaptation techniques, and comprehensive segmentation methodologies. The proposed research attempts to fill these voids, contributing to more accurate, accessible, and sustainable tomato cultivation practices through the application of advanced vision-language models combined with state-of-the-art segmentation techniques. Additionally, to the authors' knowledge, no prior studies have documented the practical deployment of VLM-based tomato leaf disease detection systems in portable device for real-world agricultural settings.

Methodology

This section outlines the experimental setup, data preprocessing, model architectures, fine-tuning strategies, and evaluation metrics employed in this study. It leverages BLIP-2 with Low-Rank Adaptation (LoRA) for efficient tomato leaf disease detection and classification, which is compared against baselines implemented with CLIP and ConvNeXT-tiny for further validation.

Figure 1 illustrated the overall workflow of Tomato Leave Disease Visual Language Model (TLDVLM). The system architecture consists of four main components: (1) image and text prompt processing, (2) data preprocessing (3) proposed TLDVLM, and (4) deployment on portable device.

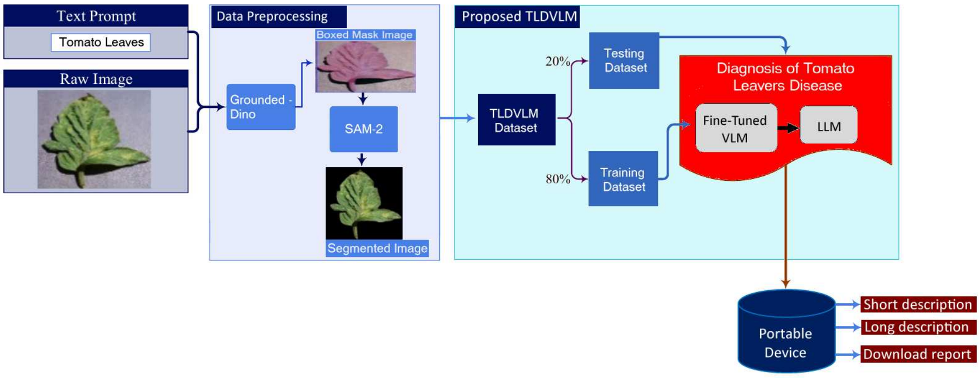

Fig. 1 **Overall workflow of Tomato Leave Disease Visual Language Model (TLDVLM).**

Image and caption dataset

The study utilized a comprehensive tomato leaf disease dataset sourced from Kaggle (Plant Village) and self capture dataset from Silal Greenhouse, Abu Dhabi. The dataset contains 10 distinct disease classes representative of common tomato pathologies. The dataset encompasses various disease conditions including Bacterial Spot, Early Blight, Late Blight, Leaf Mold, Septoria Leaf Spot, Spider Mites, Target Spot, Yellow Leaf Curl Virus, Mosaic Virus, And Healthy leaves. Each class contains varying numbers of images captured under different environmental conditions and lighting scenarios to ensure model robustness. Figure 2 illustrated the dataset class distribution. Table 1 mentioned the total amount of dataset images for training and validation of the model.

Table 1 Amount of Dataset to train and validate the model.

Dataset	PlantVillage	Own capture	Total
Dataset	13143	2000	15143

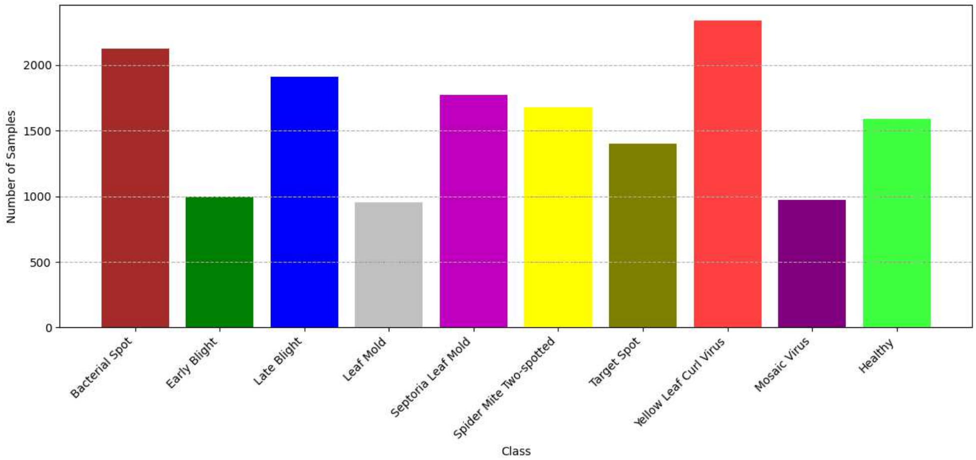

Fig. 2 **Dataset Class distribution.**

Data preprocessing pipeline

Initial image processing involves resizing all input images to a standard 224×224 pixels, followed by normalization of pixel values to the range [0,1] and standardization using ImageNet statistics. To enhance model generalization, data augmentation techniques such as random horizontal flips, rotations (±15 degrees), and color jittering are applied during training. To eliminate background noise and concentrate on the tomato leaf itself, we implement an automated leaf segmentation approach combining GroundingDINO and SAM-2. GroundingDINO acts as the object detection component, identifying and localizing tomato leaves using text queries like ”tomato leaf,” providing flexible detection, precise bounding box generation, and multi-leaf handling. The Segment Anything Model 2 (SAM-2) then utilizes these bounding boxes as prompts to generate accurate pixel-level segmentation masks, thereby enabling background elimination and focusing classification solely on the plant tissue. Finally, these segmentation masks are applied to the original images, extracting leaf-only regions by masking out or replacing non-leaf pixels with a neutral background, ensuring the classification model focuses exclusively on disease-relevant features. Figure 3 displays the segmentation outcomes of GroundedDINO and SAM-2 across 10 distinct classes. GroundedDINO and SAM-2 leverage the contextual semantic information within images to accurately extract relevant tomato leaf characteristics while eliminating background interference.

Baseline models

This section will introduce the baseline models used for comparison against our proposed BLIP-2 LoRA (TLDVLM) method. For baseline model we used CLIP-LoRA and ConvNeXT-tiny models.

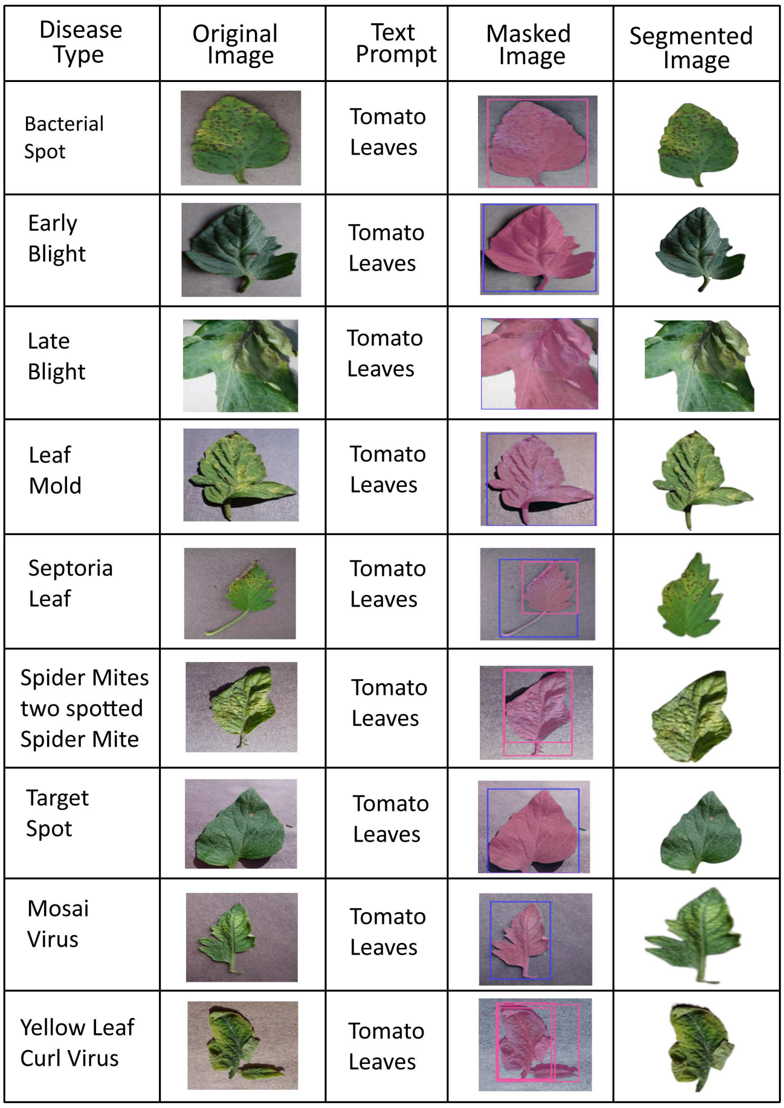

Fig. 3 **Segmentation results of GroundedDINO and SAM-2.**

- **CLIP-LoRA**: CLIP (Contrastive Language-Image Pretraining) [26] is OpenAI's vision-language model that learns to associate images and text through contrastive learning on large-scale datasets. For tomato leaf disease detection, the implementation leverages CLIP's robust vision transformer component with Low-Rank Adaptation (LoRA) for efficient fine-tuning. It loads the pre-trained CLIP model (openai/clip-vit-base-patch 16), freezes its parameters to preserve learned features, and adds a custom classifier with layer normalization and dropout for 10 disease classes. The system uses LoRA with rank 16 to reduce trainable parameters, processes the tomato dataset with 224 x 224 image resizing and normalization, and trains for 20 epochs using Adam optimizer with mixed precision on GPU. The model achieves efficient adaptation to agricultural applications by fine-tuning only a small subset of parameters while maintaining CLIP's powerful pre-trained representations. The CLIP-LoRA attained the accuracy around 91.87 %.

- **ConvNeXT-Tiny**: ConvNeXt [11], developed by Meta AI, is a modern CNN architecture that incorporates transformer-inspired design principles like hierarchical feature maps, large kernel sizes, and layer normalization to compete with vision transformers while maintaining CNN efficiency. The implementation uses ConvNeXt-tiny as a backbone for tomato leaf disease classification, loading the pre-trained facebook/convnext-tiny-224 model with mixed precision and freezing all parameters except stages 2 and 3 for task-specific adaptation. The model trains for up to 20 epochs on A100 GPU using AdamW optimizer with differential learning rates (1e-6 for backbone, 3e-5 for classifier). This approach leverages ConvNeXt's efficient architecture and pre-trained weights to achieve accurate tomato leaf disease classification with robust data augmentation and careful hyperparameter tuning. The ConvNext model achieved the accuracy around 94.57 %.

Proposed BLIP-2 LoRA method (TLDVLM)

In our study, we proposed the BLIP-2 LoRA method for classifying tomato leaf diseases using a custom dataset of 10 disease classes (Plant-Village and own data capture from Silal Greenhouse, Abu Dhabi, UAE). The implementation incorporates Low-Rank Adaptation (LoRA) to fine-tune the Q-Former module, reducing computational overhead while preserving performance.

Figure 4 illustrated the architecture of the proposed BLIP-2 LoRA model. The architecture details an image classification system for specialized tasks, such as "Disease Prediction" on a "Tomato Leaf," by integrating Low-Rank Adaptation (LoRA) into a Vision-Language Model (VLM) framework, analogous to BLIP-2. The system processes an input image, segmenting it into patches that are then fed into a frozen Vision Encoder. This encoder, typically a pre-trained Vision Transformer, extracts high-level visual features (Img') without undergoing further weight updates during fine-tuning. The core of this adapted VLM is the Q-Former, a multi-layered transformer module responsible for bridging the visual and linguistic modalities. The Q-Former contains distinct processing pathways for learned visual queries and textual prompts. Within the visual pathway, "Learned Queries" interact with the encoded image features (Img') via cross-attention mechanisms, where the queries serve as the attention keys and values, and Img' provides the query.

Crucially, LoRA is implemented within the linear projection layers of this cross-attention mechanism by introducing small, trainable low-rank matrices (LoRAQ,LoRAK,LoRAV) that are added to the frozen pre-trained weights. Similarly, the textual pathway, which processes an input prompt like "What disease on the tomato leaf?", also incorporates LoRA within its self-attention and feedforward networks, enabling efficient adaptation of the model's linguistic understanding to the

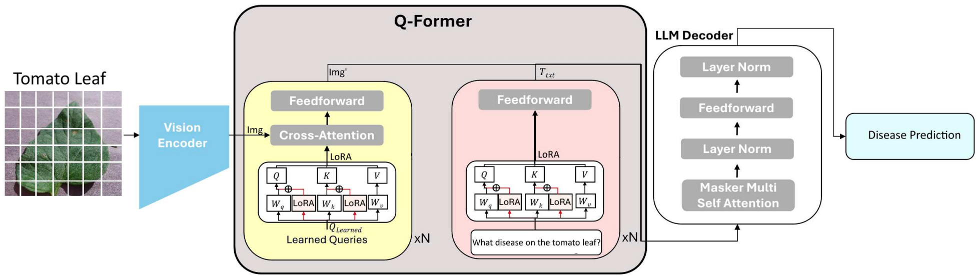

Fig. 4 Architecture of BLIP-2 LoRA model.

specific task. This selective fine-tuning of LoRA adapters significantly reduces the number of trainable parameters, thereby decreasing computational cost and memory footprint while preserving the vast pre-trained knowledge of the larger model. Finally, the refined multimodal features from the Q-Former are passed to a frozen LLM Decoder, which comprises multiple layers of masked multi-self-attention and feedforward networks. This LLM leverages its extensive language generation capabilities to produce the ultimate "Disease Prediction" based on the integrated visual and textual context provided by the LoRA-adapted Q-Former. This approach demonstrates a highly parameter-efficient strategy for specializing large VLMs for downstream image classification tasks, maintaining high performance while making fine-tuning more accessible.

Using LoRA with BLIP-2 for image classification offers substantial advantages by significantly reducing computational cost and memory footprint during fine-tuning, as it only trains a small set of low-rank matrices while freezing most pre-trained parameters [18], [10]. This efficiency not only makes fine-tuning accessible on more limited hardware but also accelerates the training process.

Training parameters

Model training was conducted on the Google Colab distributed computing platform using an NVIDIA A100 GPU with 40GB memory capacity. The optimization process employed the AdamW optimizer, with comprehensive hyperparameter configurations for the TLDVLM detailed in Table 2. The fine-tuning procedure demonstrated remarkable efficiency, completing within 5 hours while consuming less than 40 GB of VRAM. This computational efficiency underscores the TLDVLM's ability to effectively leverage limited datasets and optimize hardware utilization through the integration of Low-Rank Adaptation (LoRA) techniques. The final model size is 2.7B parameters for the base model plus 0.6 M parameters for the LoRA adapters and classifier.

Deployment on portable device

Subsequently, the proposed model was deployed on a novel portable device designed for tomato leaf disease inspection within greenhouse environments. The model was transformed into a web-based application utilizing the Gradio online user interface platform.

Table 2 TLDVLM Fine-Tuning Configuration.

Hyperparameter	Default Setting
Optimizer	AdamW
Batch Size	32
Epochs	20
Learning rate	1 e-7
Learning Schedule	Cosine Annealing

The portable device was manufactured at the Khalifa University Agrirobotics laboratory and validated in the Silal Greenhouse facility. Figure 5 illustrated the portable tomato leave disease detection device and webpage implementation of proposed TLDVLM on Gradio platform. The device includes the RGB camera, touch screen, Raspberry pi-5 processor, lithium battery, and 3D printed case.

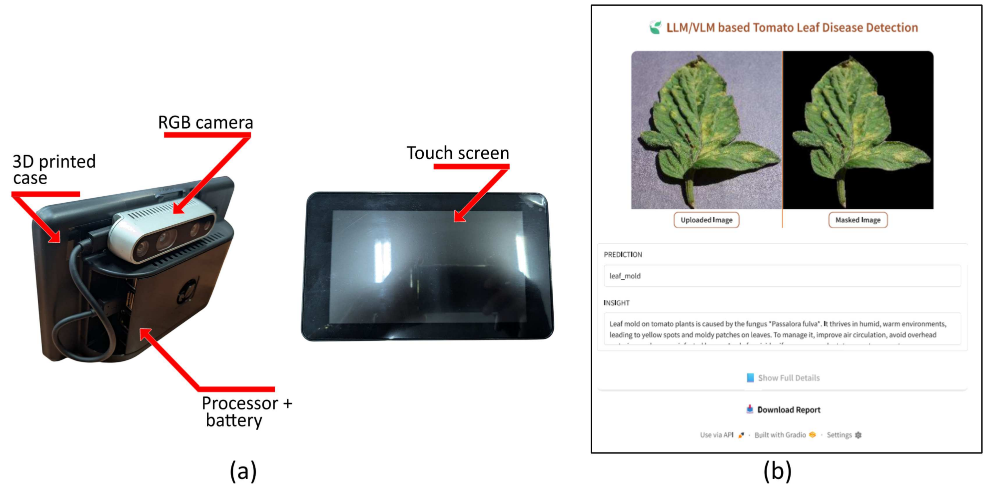

Fig. 5 Proposed TLDVLM implementation: (a) Portable device, and (b) Gradio online user interface.

The implemented system demonstrates comprehensive capabilities including segmentation, classification, and generation of both concise and detailed disease information along with corresponding treatment recommendations. Additionally, the system provides a comprehensive downloadable report in PDF format for documentation purposes. Figure 6 depicts the communication architecture of the TLDVLM device. The system captures leaf images and transmits them to a trained model server for classification analysis, subsequently receiving both concise and detailed descriptions of identified diseases along with corresponding remedial solutions. Additionally, a comprehensive report is generated to facilitate further analytical assessment.

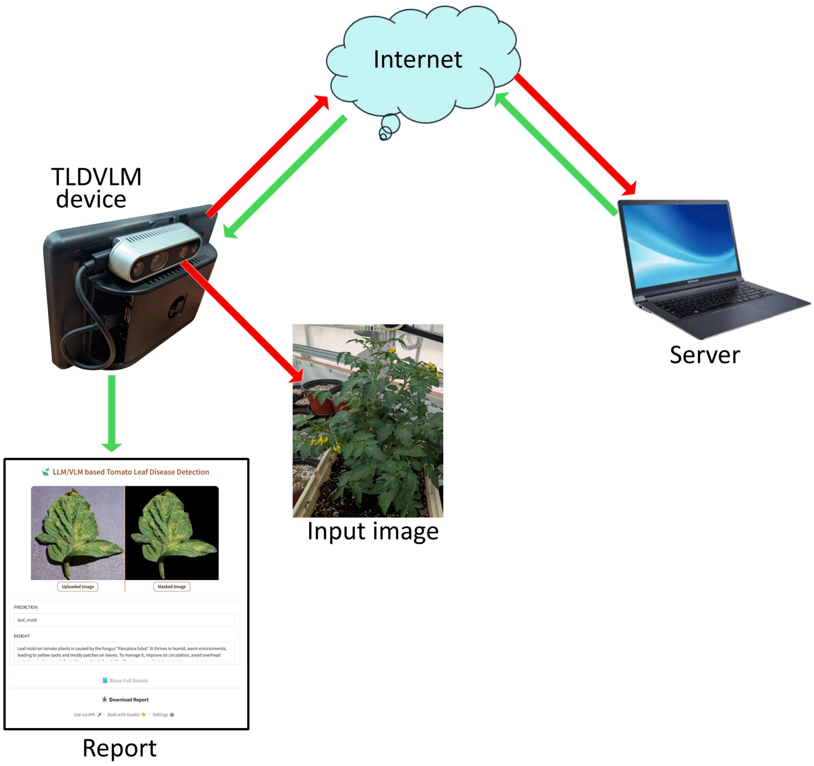

Fig. 6 **Proposed TLDVLM device communication system.**

Results

This section presents the experimental results obtained from training and evaluating the proposed BLIP-2 with LoRA model and compares its performance against other baseline models: CLIP-LoRA and ConvNeXT-tiny. We also provide a detailed analysis of model convergence, class-specific performance, and overall robustness indicators. These results indicate that the model has successfully learned to distinguish between the 10 tomato disease classes with high accuracy while maintaining good generalization capabilities.

Overall Performance Comparison

Table 3 shows the key performance metrics of proposed method with baseline models. As depicted from the Table 3 the BLIP-2 LoRA (proposed) model outperformed than others baseline models for detection of tomato leave diseases. The model demonstrated stable convergence, with a final training loss of 0.6068 and a validation loss of 0.6970 upon completion at 20 epoch. The small loss differential of 0.0902 suggests good generalization with very minimal overfitting. Performance metrics were robust:

a high precision of 0.9587 indicates excellent accuracy in positive predictions, crucial for avoiding unnecessary treatments. An exceptional recall of 0.9789 highlights the model's ability to identify nearly all actual disease instances, thereby minimizing missed diagnoses. The balanced F1-score of 0.9681 further confirms the model's strong overall performance across all disease classes.

Table 3 Key Performance metrics.

Models	Accuracy	Precision	Recall	F1-Score
CLIP-LoRA	91.87%	0.9046	0.8896	0.8957
ConvNeXt-tiny	94.57%	0.9270	0.9463	0.9360
BLIP-2-LoRA (Proposed)	97.27%	0.9587	0.9789	0.9681

Confusion Matrix Analysis

The confusion matrix reveals detailed insights into the model's classification performance across all disease classes. Figure 7 illustrated the confusion matrix's for CLIP-LORA, ConvNeXT-Tiny, and BLIP-2-LoRA. Figure 7 (a) presented the heatmap for CLIP-LoRA baseline model across the 10 tomato leaf disease classes. The off-diagonal elements show a higher frequency of misclassifications compared to the BLIP-2 LoRA model (Figure 7 (c). For instance, class 1 (Early Blight) exhibits multiple misclassifications into other classes (e.g., 18 instances predicted as class 0, 22 as class 2, 25 as class 3), suggesting a less refined understanding of its visual features compared to TLDVLM. Class 6 also shows more dispersed errors. The overall lighter shades on the diagonal for several classes, along with more pronounced off-diagonal values, indicate that CLIP-LoRA's discriminatory power is somewhat less precise than that of the BLIP-2 LoRA model for this specific dataset and task. Similarly, Figure 7 (b) represents the heatmap for ConvNeXT-Tiny baseline model. ConvNeXT-tiny generally performs well, the misclassifications patterns, indicated by off-diagonal values, are more prominent than those observed with BLIP-2 LoRA. For example, similar to CLIP-LoRA, there are noticeable misclassifications for several classes, such as class 1 (Early Blight) having 18 instances misclassified as class 0, and others scattering across different labels. Class 6 also demonstrates a notable number of incorrect predictions. The visual comparison with Figure 7 (c) highlights that while ConvNeXT-tiny, as a robust CNN, achieves good performance, it occasionally struggles with finer distinctions between disease types, leading to a higher incidence of misclassifications compared to the multi-modal TLDVLM.

Figure 7 (c) illustrated the heatmap of BLIP-2 LoRA model. Notably, classes such as 1 (Early Blight), 4 (Leaf Mold), and 8 (Target Spot) show perfectly or near-perfectly classified instances with no or minimal off-diagonal values, confirming their highly distinctive visual features. Misclassifications, represented by non-zero off-diagonal entries, are generally low. The matrix reveals that the majority of errors occur between visually similar classes, with some instances of class 7 (Spider Mite Two-spotted) being

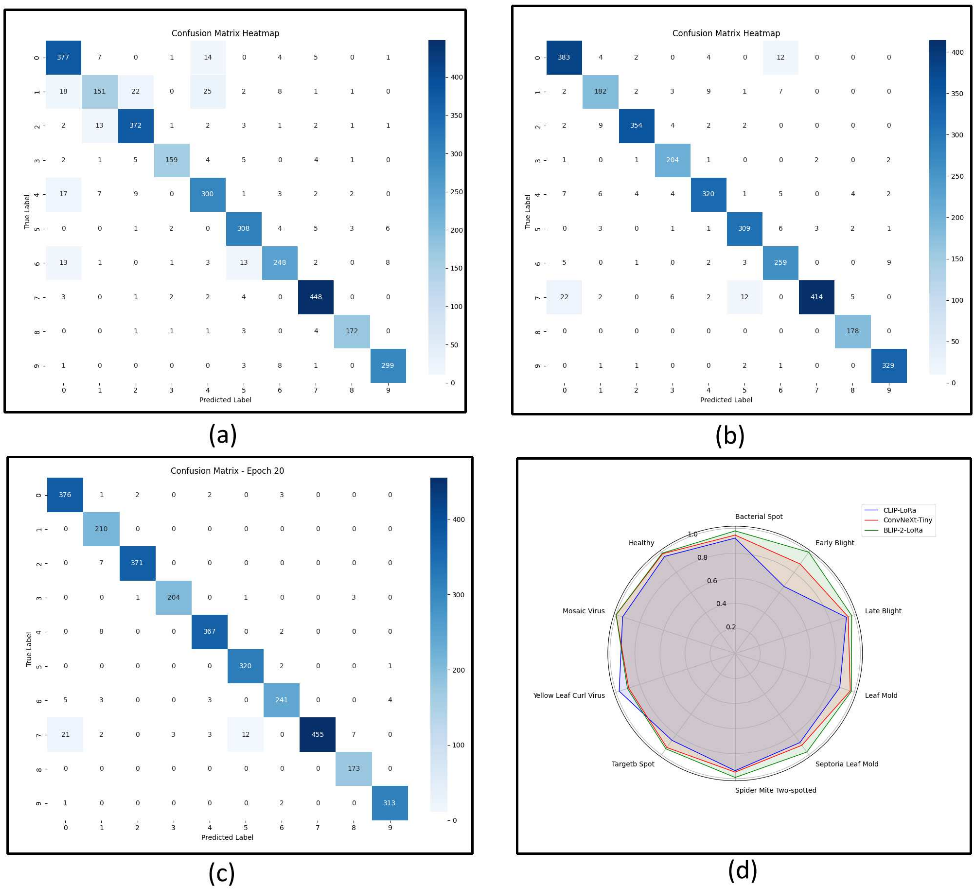

Fig. 7 Confusion matrix: (a) CLIP-LoRA, (b) ConvNeXt-Tiny, (c) BLIP-2 LoRA, and (d) Spider graph.

misclassified across several other classes, suggesting subtle visual overlaps. The overall strong diagonal presence across all classes underscores the TLDVLM's effective discrimination capabilities.

Figure 7 (d) provides a comparative visualization of the F1-scores for each of the 10 tomato leaf disease classes across the three models: CLIP-LoRA (blue line), ConvNeXT-tiny (red line), and BLIP-2 LoRA (green line). Each spoke of the radar represents a specific disease class, with the F1-score ranging from 0 at the center to 1 at the outermost ring. The chart clearly illustrates that the BLIP-2 LoRA model (green line) consistently achieves the highest F1-scores across nearly all disease classes, encompassing a larger area within the radar plot. While ConvNeXT-tiny (red line) shows competitive performance on some classes, and CLIP-LoRA (blue line) performs reasonably, BLIP-2 LoRA's green envelope consistently extends further towards the maximum F1-score of 1.0. This visual representation powerfully reinforces that the proposed TLDVLM with BLIP-2 LoRA demonstrates superior and more balanced classification performance across the entire spectrum of tomato leaf diseases compared to the other two baseline methods.

Evaluation metrics and performance assessment

Figure 8 illustrated the evaluation metrics of proposed (BLIP-2 LoRA) model for VLM based tomato leave disease classification and reasoning.

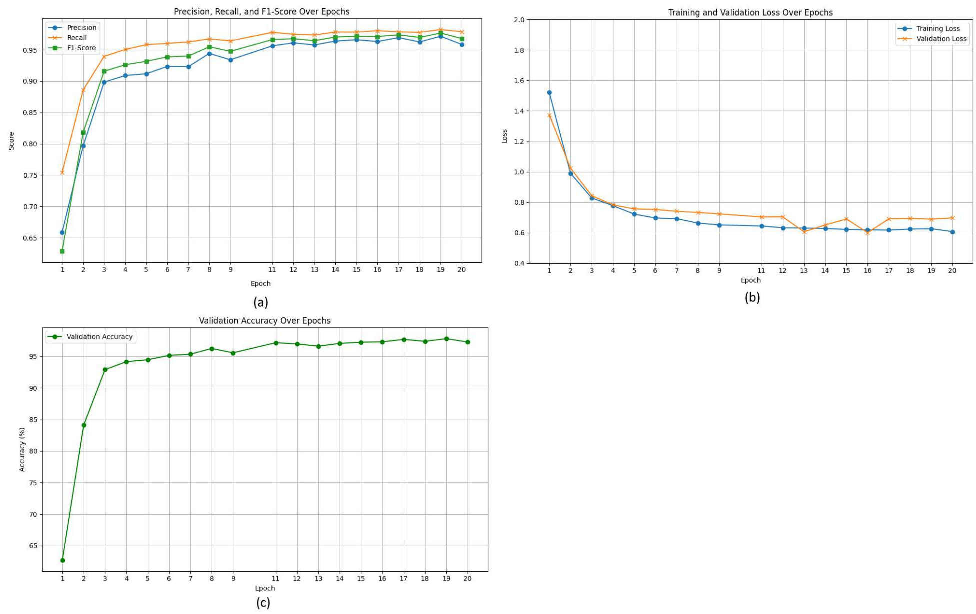

Fig. 8 Evaluation metrics for BLIP-2 LoRA : (a) Precision, recall, F1 score, (b) Training and validation loss, and (c) Validation accuracy.

Figure 8 (a), displays the evolution of key classification metrics—Precision, Recall, and F1-Score—for the TLDVLM (BLIP-2 LoRA) model during training over 20 epochs. The x-axis represents the epoch number, and the y-axis denotes the score ranging from 0.6 to 1.0. All three metrics (Precision - blue circles, Recall - orange 'x' markers, F1-Score - green squares) show a rapid increase during the initial epochs, signifying quick improvements in classification performance. After roughly epoch 6, the scores begin to stabilize and continue to gradually improve, reaching high plateaus. By epoch 20, the Precision, Recall, and F1-Score converge to approximately 0.9587, 0.9789, and 0.9681, respectively. The consistently high and closely aligned values across these metrics highlight the model's balanced and robust performance in correctly identifying positive instances and minimizing both false positives and false negatives. Figure 8 (b), presents the progression of the training loss and validation loss for the TLDVLM (BLIP-2 LoRA) model over 20 epochs. Both the training loss (blue line with circular markers) and validation loss (orange line with 'x' markers) demonstrate a sharp decrease during the initial epochs, indicating rapid learning. After approximately 4-5 epochs, the decrease in both curves stabilizes, suggesting convergence. The final training loss is observed to be around 0.6068, while the validation loss hovers around 0.6970. The consistently small gap between the training and validation loss curves throughout the training process indicates that the model is generalizing well to unseen data and is not significantly overfitting. This stable convergence confirms the effectiveness of the training configuration and regularization techniques employed. Figure 8 (c), tracks the Validation Accuracy for the TLDVLM (BLIP-2 LoRA) model throughout its 20 epochs of training. The x-axis denotes the epoch number, and the y-axis represents the accuracy percentage. The curve shows a steep increase in accuracy during the initial few epochs, rising from approximately 63% at Epoch 1 to over 90% by Epoch 3. The accuracy continues to climb steadily, eventually reaching a plateau above 95% around Epoch 8-10. From Epoch 10 onward, the validation accuracy remains consistently high, hovering between 96% and 97%, with minor fluctuations. This stable and high validation accuracy confirms the model's robust learning and strong generalization capabilities on unseen data, indicating successful fine-tuning for the tomato leaf disease classification task.

Table 4 presents a comparative analysis of the proposed model against existing traditional approaches specifically designed for tomato leaf disease detection. The results demonstrate that while the Attention-based residual CNN model achieves a marginally higher accuracy of 98% compared to our proposed approach, it exhibits significant limitations in reasoning capabilities and lacks mobile/device deployment feasibility. In contrast, our proposed model attains 97.27% accuracy while providing enhanced functionality through disease classification and the generation of precise remedial recommendations, making it more suitable for practical agricultural applications.

Experimentation of TLDVLM based portable device

The validation of the TLDVLM model on portable device was carried out at greenhouse setting for tomato crop. The device is tested on tomato leaves to capture the images and then used as input for TLDVLM to analysis and generate the report. The final report is generated as PDF format which is further used for disease analysis and plant health monitoring. Figure 9 shows the experimentation at greenhouse test field.

Conclusion

This study successfully developed and evaluated the Tomato Leaf Disease Visual Language Model (TLDVLM), a robust and efficient solution for the precise classification of 10 distinct tomato leaf diseases. By leveraging the advanced multimodal architecture of BLIP-2 and implementing Low-Rank Adaptation (LoRA) on its Q-Former, the TLDVLM achieved superior performance compared to traditional computer vision models like ConvNeXT-tiny and other parameter-efficient VLM adaptations such as CLIP-LoRA. The integration of GroundingDINO and SAM-2 for automated leaf segmentation proved crucial in enhancing the model's focus and accuracy by eliminating irrelevant background noise.

Table 4 State-of-art comparison.

Models	Reasoning Capability	No. of Classes	Dataset	Accuracy	Mobile/Device Deployment	Reference
LeNet	No	10	PlantVillage	94-95 %	No	[31]
AlexNet	No	10	PlantVillage	94-95 %	No	[Durmus et al]
Attention based residual CNN	No	98	PlantVillage	98 %	No	[karthik et al]
AlexNet, SqueezeNet 1.1	No	10	PlantVillage + Self	96.3 %	Yes	[Nag et al]
MobileNets	No	10	Self	90.3 %	Yes	[Elhassouny]
DCNN	No	6	Self	76 %	No	[Foysal]
BLIP-2 - LoRA	Yes	10	PlantVillage + Self	97.27 %	Yes	Our Proposed approach

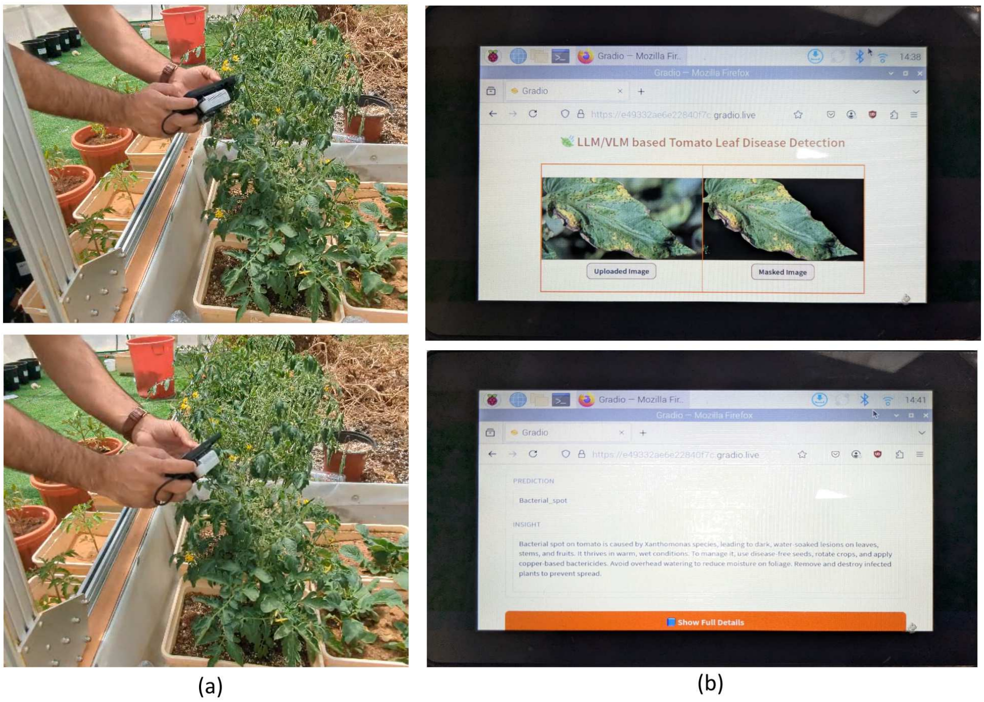

Fig. 9 Proposed TLDVLM implementation: (a) Portable device testing at greenhouse, and (b) Device screen results.

Our experimental results unequivocally demonstrate the efficacy of the TLD-VLM, with an impressive overall accuracy of 97.27%, complemented by high precision (0.9587), recall (0.9789), and F1-score (0.9681). This performance underscores the capacity of LoRA-adapted Vision-Language Models to retain powerful pre-trained knowledge while being efficiently fine-tuned for specific, high-stakes tasks like agricultural disease detection. The detailed analysis of class-specific performance revealed the model's strong discriminatory abilities, with perfect classification on several disease types and effective handling of class imbalance through weighted sampling and loss functions.

Beyond its classification prowess, the TLDVLM's practical utility is further enhanced by its deployment within a portable application. This tool provides instant visual feedback (raw and segmented images), disease predictions, and, critically, offers on-demand access to comprehensive information regarding disease causes and remedies via external knowledge bases (e.g., OpenAI). The ability to download a PDF summary for offline use signifies a significant step towards accessible and actionable agricultural diagnostics for farmers and growers.

In conclusion, the TLDVLM represents a promising advancement in intelligent agricultural systems, showcasing that state-of-the-art vision-language models, when combined with parameter-efficient fine-tuning techniques and smart integration, can yield highly accurate, robust, and deployable solutions for critical challenges like crop disease management. Future work will focus on expanding the model's capabilities to detect multiple diseases on a single leaf, real-time inference optimization, and broader validation across diverse environmental conditions.

Data availability

Data Availability Statement: The datasets used during the current study are available from the corresponding author on reasonable request.

Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

- **Ethics approval:** This study did not involve human participants or animals, thus no ethics approval or consent to participate was required.
- **Institutional review board:** Not applicable.
- **Clinical trial number:** Not applicable.
- **Informed consent:** Not applicable.
- **Conflicts of interest:** The authors declare no competing interests.
- **Competing interest:** The authors declare no competing interests.
- **Funding:** Not applicable.
- **Authors' contributions:** M.K: As the author, responsible for the main aspects of the research, including model design, data collection, analysis and interpretation, as well as drafting the initial manuscript. R. S: as the equal author contribution and responsible for methodology, experimentation, Implementation and result analysis, Reviewing and Editing the draft. A. S: guidance of the research, team coordination, resources, final review. I.H: As the corresponding author, funding allocation, overall guidance, and correspondence with the journal.
- **Acknowledgments:** "This work is supported by Khalifa University Center for Autonomous Robotic Systems (KUCARS) under Award RC1-2018-KUCARS Theme 4."

References

[1] Abu-Alrub, I., Saleh, S., Awaga, A.: Effect of different rates of nitrogen and phosphorus fertilizers on yield and quality of greenhouse tomato under the uae condition. Ec Agric **5**, 139–146 (2019)

[2] Aversano, L., Bernardi, M.L., Cimitile, M., Iammarino, M., Rondinella, S.: Tomato diseases classification based on vgg and transfer learning. In: 2020 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), pp. 129–133 (2020). IEEE

[3] Anam, S., Fitriah, Z.: Early blight disease segmentation on tomato plant using k-means algorithm with swarm intelligence-based algorithm. International Journal of Mathematics and Computer Science **16**(4), 1217–28 (2021)

[4] Abbas, A., Jain, S., Gour, M., Vankudothu, S.: Tomato plant disease detection using transfer learning with c-gan synthetic images. Computers and electronics in agriculture **187**, 106279 (2021)

[5] Arshad, M.A., Jubery, T.Z., Roy, T., Nassiri, R., Singh, A.K., Singh, A., Hegde, C., Ganapathysubramanian, B., Balu, A., Krishnamurthy, A., *et al.*: Leveraging vision language models for specialized agricultural tasks. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 6320–6329 (2025). IEEE

[6] Agarwal, M., Singh, A., Arjaria, S., Sinha, A., Gupta, S.: Toled: Tomato leaf disease detection using convolution neural network. Procedia Computer Science **167**, 293–301 (2020)

[7] Brahimi, M., Boukhalfa, K., Moussaoui, A.: Deep learning for tomato diseases: classification and symptoms visualization. Applied Artificial Intelligence **31**(4), 299–315 (2017)

[8] Buajaila, F., Cowan, J.S., Inglis, D., Carpenter-Boggs, L., Miles, C.: Tomato growth, yield, and quality response to mixed chemical–organic fertilizers and grafting treatments in a high tunnel environment. Canadian Journal of Plant Science **102**(1), 33–48 (2021)

[9] Fuentes, A., Yoon, S., Kim, S.C., Park, D.S.: A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. Sensors **17**(9), 2022 (2017)

[10] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., *et al.*: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)

[11] KP, A.R., Gowrishankar, S.: Convnext-based mango leaf disease detection: Differentiating pathogens and pests for improved accuracy. International Journal of Advanced Computer Science and Applications **14**(6) (2023)

[12] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., *et al.*: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)

[13] Khirade, S.D., Patil, A.B.: Plant disease detection using image processing. In: 2015 International Conference on Computing Communication Control and Automation, pp. 768–771 (2015). IEEE

[14] Kaur, S., Pandey, S., Goel, S.: Plants disease identification and classification through leaf images: A survey. Archives of Computational Methods in Engineering **26**, 507–530 (2019)

[15] Kawasaki, Y., Uga, H., Kagiwada, S., Iyatomi, H.: Basic study of automated diagnosis of viral plant diseases using convolutional neural networks. In: Advances in Visual Computing: 11th International Symposium, ISVC 2015, Las Vegas, NV, USA, December 14-16, 2015, Proceedings, Part II 11, pp. 638–645 (2015). Springer

[16] Kuska, M.T., Wahabzada, M., Paulus, S.: Ai-chatbots for agriculture-where can large language models provide substantial value? Available at SSRN 4685971 (2024)

[17] Laterrot, H.: Disease resistance in tomato: practical situation. Acta Physiologiae Plantarum **22**, 328–331 (2000)

[18] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, pp. 19730–19742 (2023). PMLR

[19] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900 (2022). PMLR

[20] Liu, J., Wang, X.: Tomato diseases and pests detection based on improved yolo v3 convolutional neural network. Frontiers in plant science **11**, 898 (2020)

[21] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., *et al.*: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision, pp. 38–55 (2024). Springer

[22] Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Frontiers in plant science **7**, 215232 (2016)

[23] Mohamed, M.: Agricultural sustainability in the age of deep learning: current trends, challenges, and future trajectories. Sustainable Machine Intelligence Journal **4**, 2–1 (2023)

[24] Martinelli, F., Scalenghe, R., Davino, S., Panno, S., Scuderi, G., Ruisi, P., Villa, P., Stroppiana, D., Boschetti, M., Goulart, L.R., *et al.*: Advanced methods of plant disease detection. a review. Agronomy for sustainable development **35**, 1–25 (2015)

[25] Ngugi, L.C., Abdelwahab, M., Abo-Zahhad, M.: Tomato leaf segmentation algorithms for mobile phone applications using deep learning. Computers and

Electronics in Agriculture **178**, 105788 (2020)

[26] Quoc, K.N., Thu, L.L.T., Quach, L.-D.: A vision-language foundation model for leaf disease identification. arXiv preprint arXiv:2505.07019 (2025)

[27] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)

[28] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PmLR

[29] Selvaraj, M.G., Vergara, A., Ruiz, H., Safari, N., Elayabalan, S., Ocimati, W., Blomme, G.: Ai-powered banana diseases and pest detection. Plant methods **15**, 1–11 (2019)

[30] Tzachor, A., Devare, M., Richards, C., Pypers, P., Ghosh, A., Koo, J., Johal, S., King, B.: Large language models and agricultural extension services. Nature food **4**(11), 941–948 (2023)

[31] Tm, P., Pranathi, A., SaiAshritha, K., Chittaragi, N.B., Koolagudi, S.G.: Tomato leaf disease detection using convolutional neural networks. In: 2018 Eleventh International Conference on Contemporary Computing (IC3), pp. 1–5 (2018). IEEE

[32] Din, Z., Adnan, S.M., Ahmad, R.W., Aziz, S., Ismail, W., Iqbal, J.: Classification of tomato plants' leaf diseases using image segmentation and svm. Technical Journal **23**(02), 81–88 (2018)

[33] Wang, Y., Zhang, P., Tian, S.: Tomato leaf disease detection based on attention mechanism and multi-scale feature fusion. Frontiers in Plant Science **15**, 1382802 (2024)

[34] Yang, X., Gao, J., Xue, W., Alexandersson, E.: Pllama: An open-source large language model for plant science. arXiv preprint arXiv:2401.01600 (2024)

[35] Zhang, K., Wu, Q., Liu, A., Meng, X.: Can deep learning identify tomato leaf disease? Advances in multimedia **2018**(1), 6710865 (2018)

[36] Zhou, Y., Yan, H., Ding, K., Cai, T., Zhang, Y.: Few-shot image classification of crop diseases based on vision–language models. Sensors **24**(18), 6109 (2024)