arXiv:2608.08727v1 [cs.CV] 9 Aug 2026

TomaMMU: A Comprehensive Multimodal Understanding Benchmark for Tomato Leaf Diseases

Gia-Han Truong[1⋆], Khang Nguyen Quoc[2⋆], and Luyl-Da Quach[1✉]

[1] Department of Information Technology, FPT University, Can Tho, Vietnam
[2] School of Electrical Engineering, Korea University, Seoul, South Korea
truonggiahan0702ct@gmail.com, khangnq@korea.ac.kr, luyldaquach@gmail.com

Abstract. Despite advances in automated plant disease recognition, existing systems primarily frame diagnosis as a visual classification problem, leaving multimodal disease understanding and reasoning underexplored, particularly in real-world settings marked by visual variability, background clutter, and inconsistent illumination. To address this gap, we introduce TomaMMU, a large-scale **Toma**to leaf disease **M**ulti**M**odal **U**nderstanding dataset, alongside TomaBench, a benchmark for evaluating Vision–Language Models (VLMs) on tomato disease understanding. TomaMMU comprises 28,808 high-quality images spanning 15 categories and 213,119 human-annotated visual question-answer pairs, generated through a three-stage pipeline comprising Data Collection, Human Annotation, and Question-Answer Generation. Building on this foundation, TomaBench organizes seven agricultural tasks into a hierarchical three-level taxonomy spanning Basic Perception, Pathology Understanding, and Expert Diagnosis, which together enable systematic evaluation from low-level visual recognition to high-level diagnostic reasoning. The tasks assess visual symptom recognition, taxonomic relationships, and diagnostic reasoning, offering a comprehensive view of how well models grasp plant pathology. Our results pronounced gaps in fine-grained recognition and factually grounded reasoning with 14 state-of-the-art VLMs, consistently underperforming on both challenging Multiple-Choice Questions (MCQs) and open-ended questions. These results suggest that current VLMs struggle to translate visual perception into reliable diagnostic knowledge, motivating the need for targeted domain adaptation. Simple fine-tuning on TomaMMU substantially narrows this gap, boosting accuracy on challenging MCQs to 96.09%, outperforming recent VLMs, and pointing toward promising directions for future work. All data and code is available in https://huggingface.co/datasets/enalis/TomaMMU

Keywords: Vision-Language Models · Multimodal Dataset · Visual Question Answering · Multimodal Benchmark · Agricultural decision support

⋆ Equal contribution

1 Introduction

Multimodal understanding and reasoning are crucial for plant disease identification, helping the system overcome environmental influences in decision-making and also providing farmers and experts with a more intuitive understanding of plant diseases. Survey studies have shown that many studies have used various machine learning methods, from conventional vision models to advanced VLMs, following a process of data collection, preprocessing, segmentation, feature extraction, classification, and evaluation, but are limited by environmental factors, interpretability, lack of field data, and especially generalization (where a model performs well on a tomato dataset but not on other data types) [13,14,31]. Hence, this study proposes a new process for building a large-scale, standardized dataset to provide detailed assessments of plant diseases, support understanding, and enable high-level inferences, including assessments of tomatoes, thereby enabling further research to be expanded and developed in the future.

Recent advances in smart agricultural adaptation, such as Artificial Intelligence (AI), the Internet of Things (IoT), and Deep Learning (DL), are achieving promising results in addressing the large volume of data collected from cameras, satellites, and drones. George et al. surveyed many studies that applied IoT, Explainable AI (XAI), Convolutional Neural Networks (CNNs), and Vision Transformers (ViTs) to benchmark datasets (PlantVillage [19]). However, all encountered difficulties in real-world conditions [6], with accuracy typically dropping by 30-40% when the model was trained on a laboratory dataset (PlantDoc [35], FieldPlant [20]). To clarify model-dependent classification, studies have been conducted using XAI to explain the dependence between feature-related outcomes and advanced DL models to address dependence on field outcomes such as background removal, focusing on CNN model changes, but also depending on real-world images, background clutter, occlusions, leaf pose variation, and natural symptom diversity still impacting the results [26, 34]. These factors, such as lighting, background clutter, occlusions, leaf pose variation, and natural symptom diversity, reveal that many models rely on superficial correlations rather than robust disease representations, leading to poor generalization in field environments. The stakes are also high in this scenario, as it directly impacts food security and farmer livelihoods. This performance degradation is due to the lack of comprehensive benchmarks to evaluate and improve VLMs' ability to understand agricultural problems, especially to address real farmers and experts concerns.

Creating such a benchmark presents various multifaceted challenges. First, biological and agricultural tasks, which often use computer vision, are notably labor-intensive for data collection (labeling is a first-order requirement) [14]. However, such domain knowledge is scarce and highly specialized, making it exceedingly difficult to curate high-quality evaluation datasets. Second, to address environmental variability, robust systems must go beyond simple image classification by combining the visual observation with background information [49]. Third, no clear protocol for defining a representative distribution of realistic

agricultural questions [41]. These challenges have created a significant gap in understanding whether AI systems can handle real-world agricultural problems.

To address the aforementioned challenges, we introduce the Tomato leaf disease MultiModal Understanding Tomato leaf disease MultiModal Understanding (TomaMMU) dataset, designed to support the development of VLMs for research on tomato leaf disease. Unlike other datasets that rely on generating images with a uniform laboratory background and synthetic annotations via AI, TomaMMU integrates curated human-label annotations. It also includes images collected from other authoritative sources, such as LeafNet [27], enabling evaluation under real-world conditions rather than solely recognizing surface patterns in a controlled environment. For data curation, we design a pipeline to extract tomato leaf knowledge from long-form expert answers, and then have human annotators evaluate the quality and finalize the question-answer pairs. Based on the real-world questions we collected, the final dataset covers seven major types of tomato questions and knowledge: crop species identification, healthy-diseased classification, single-many leaves classification, disease classification, symptom identification, pathogen classification, and scientific name classification, as shown in Fig. 1. Including with TomaMMU is TomaBench, an assessment framework with 42,626 question-answer pairs covering 3 hierarchical diagnostic tasks (from basic health screening to in-depth scientific name classification), a comprehensive evaluation resource designed to assess the progress of VLMs in understanding tomato pathology. The proposed dataset and evaluation framework are expected to significantly accelerate research progress in understanding multimodal tomato disease.

Overall, the main contributions of this study are as follows:

- We provide TomaMMU, a VQA dataset for tomato leaf disease comprising more than 28,808 images and up to 213,119 VQA pairs across 14 tomato diseases and 1 tomato healthy, along with seven VQA tasks, which are Healthy Diseases Classification (HDC), Disease Classification (DC), Pathogen Classification (PC), Crop species identification (CSI), Scientific Name Classification (SNC), Leaf Count (LC) and Symptoms Identification (SI). These tasks test the extent to which VLMs can support disease diagnosis by integrating image information into real-world environments.
- Introducing TomaBench, a comprehensive benchmarking framework for tomato disease detection and characterization.
- A systematic evaluation of the leading VLMs reveals their limitations in handling knowledge-intensive queries about tomato diseases, along with an error analysis.

2 Related Work

Multimodal Datasets and Benchmarks. General-domain multimodal benchmarks have evolved from basic visual question answering toward complex, real-world reasoning. These datasets, such as VQAv2 [8] and GQA [12] established the foundation for visual question answering, while ScienceQA [30] introduced

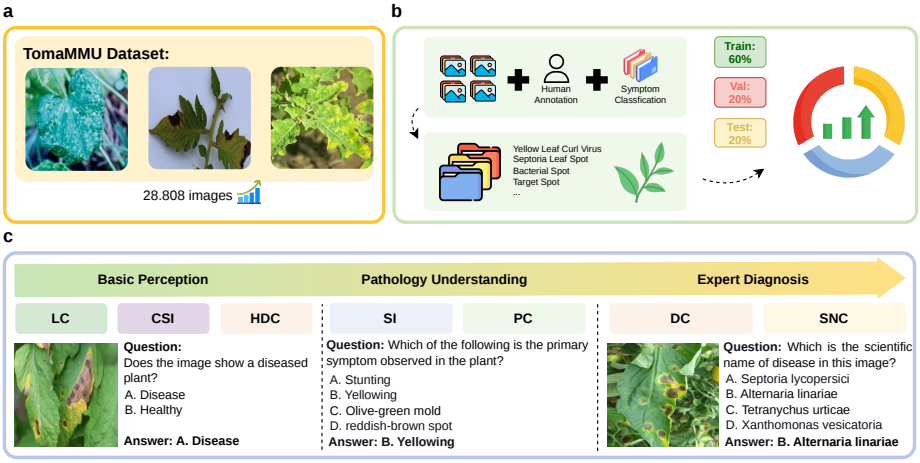


Fig. 1: The TomaMMU Curation and Benchmarking Pipeline. **a,** Raw image data of tomato leaf types were collected from in-the-wild sources and in-house collections. **b,** Humans participated in manual labeling and classification of disease symptoms. A metadata set with Train/Val/Test sets proportionally divided into 60% / 20% / 20% portions. **c,** Question and Answer Dataset. From metadata-image pairs, we developer TomaBench, a curated subset designed to benchmark Large Vision-Language Models on image classification, zero-shot learning, and Visual Question Answering (VQA).

scientific reasoning, though its academic-source samples limit real-world relevance. More recent benchmarks push further into complex reasoning and real-world grounding: RealWorldQA [45] targets everyday visual queries, and Eyes-Wide-Shut probes whether models rely on genuine visual grounding rather than language priors [39]. Within the agriculture domain, multimodal datasets remain comparatively scarce and narrow in scope, often sacrificing scale, realism, or reasoning depth. Larger-scale efforts have since emerged to address the coverage gap, such as the Crop Disease Domain Multimodal Dataset (CDDM) [17] dataset, which has expanded to 137,000 images spanning 16 crop species and 60 disease categories, paired with nearly one million question-answer pairs. Beyond scale, other benchmarks push toward richer reasoning: AgriBench [48], and LeafBench [27] organizes agricultural questions into a hierarchical structure that evaluates both surface-level understanding and deeper reasoning designed for vision-language understanding of plant diseases.

Even with these advances, however, no existing agricultural dataset combines large-scale, field-collected imagery with the depth of hierarchical reasoning needed to fully capture real-world diagnostic complexity, a gap that TomaMMU is designed to close.

Vision-language models for plant disease. Adapting these general-purpose models to agriculture has proven considerably harder, since agricultural imagery involves fine-grained morphological cues that general VLMs are not trained to distinguish. This challenge has driven a progression of increasingly specialized

adaptations, beginning with BioCLIP, which extends the CLIP framework to biological taxonomy and offers a foundation model for the tree of life that partially transfers to plant identification [38]. Building on this taxonomic grounding, AgriCLIP narrows the focus further through domain-specialized cross-modal alignment tailored to agriculture and livestock applications [22], and SCOLD sharpens this specialization even more by refining vision-language alignment specifically for leaf disease identification [28]. Together, these efforts demonstrate measurable gains over general-purpose VLMs as domain specificity increases, yet they remain anchored in classification-style supervision rather than open-ended diagnostic reasoning. This persistent reliance on classification underscores the need for benchmarks like TomaBench, which test deeper pathological understanding rather than surface-level recognition.

3 Methodology

3.1 TomaMMU Overview

We built TomaMMU, short for "Tomato leaf diseases MultiModal Understanding", a comprehensive dataset and a multimodal benchmark to evaluate the ability of VLMs in understanding multiple tasks related to the phenotype and pathology of tomatoes. Two critical design choices were adopted for the data curation process: **Real-world distribution** images, questions, and answers are derived from real diagnoses and consulted by agricultural experts, and **Factual question** we employ Open-Ended Questions (OEQs), providing the questions and requiring VLMs to generate short-form responses that directly mention the key knowledge helpful to the user. In addition, we provided multiple-choice questions (MCQs) to align with most multimodal benchmarks. This combination released a comprehensive evaluation.

Data Description We designed a dataset-processing pipeline for human verification to convert biological and pathological information into high-quality question-answer (QA) pairs for agricultural knowledge evaluation. After filtering and processing, our TomaMMU comprises: (1) a training dataset with 124,367 VQA pairs specialized for fine-tuning VLMs and (2) an evaluation set with 42,626 MCQs and 37,816 OEQs covering the knowledge types of tomato leaves called TomaBench. The large number of MCQs stems from our data construction strategy: each symptom in an image generates a separate MCQs with four options, while all symptoms in an image are merged into a single OEQs to produce a short answer.

3.2 Dataset Curation

To construct a robust multimodal question-answering model for tomato disease diagnosis, we developed the TomaMMU dataset, comprising 28,808 with image size is 224×224 px images across 14 disease categories and one healthy category. To advance, we built TomaBench, a task-specific benchmark designed to evaluate

the robustness of VLMs in a zero-shot setting, which means the models are not fine-tuned on the QA pairs. As depicted in Fig. 1, we outline a three-stage pipeline aspect follows:

- **Stage 1: Data Collection** The image data for this study consists of two parts: in-the-wild sources and in-house collections. These data were obtained from LeafNet [27] and TOM2024 [3], as shown in Fig. 3b. Containing over 28,808 diseased tomato leaves representing 14 diseases and 25 tomato disease symptoms, TomaMMU offers significantly greater scale and diversity than existing datasets. Previous studies relied predominantly on laboratory-acquired images characterized by uniform backgrounds. Although these models perform well on the internal dataset, in an uncontrolled environment, they often experience significant performance drops on out-of-distribution samples. To mitigate this generalization gap, TomaMMU emphasizes in-situ data acquisition, with images captured directly from farms comprising the majority of the dataset.
- **Stage 2: Human Annotation and Categorization** In Fig. 1b illustrates the diversity of diseases represented in TomaMMU. From a pathological angle, the dataset encompasses a broad biological range; "*Yellow Leaf Curl Virus*" and "*Septoria Leaf Spot*" diseases dominate. This coarse-grained categorization is further refined through detailed taxonomic classification. To distinguish TomaMMU from existing datasets, we implement an exhaustive annotation protocol. Instead of a simple static label, we extend the standard labels with structured metadata that includes detailed symptom descriptions, disease type, and quantity. To ensure the quality of the evaluation and precision of the ground-truth labels, mislabeled data is removed to ensure symptoms are visually identifiable. A human domain expert strictly executes the entire metadata extraction and annotation pipeline. Every data point undergoes a detailed manual verification process in which experts carefully filter and extract diagnostic knowledge, thereby ensuring the dataset's high accuracy and agronomic reliability. Through collecting and annotating, we harvest a dataset comprising 213,119 conversations. This distribution ensures that the data across the different categories is in general equilibrium.

To assist VLMs in handling complexity and a wide variety of real-world agricultural questions. We tag each question with an agriculture sub-domain. Guided by PlantVillageVQA's [32] seasoned approach, we decided on the three categories "Basic Perception", "Pathology Understanding", and "Expert Diagnosis", as shown in the Fig. 1c. The model will open with simple tasks, then move on to detailed level 2 analysis and verification, focusing on detecting symptoms and pathogens, and finally to higher-level inference and judgment, handling more complex questions. This part requires the model to synthesize knowledge and reason to answer. This design mitigates the hallucination problem typical of VLMs. Then, to systematically extract knowledge, we organize agricultural knowledge into categories of Healthy-Diseased Classification (HDC), Leaf Classification (LC), Disease Classification (DC), Crop Species Identification (CSI),

Scientific Name Classification (SNC), Pathogen Classification (PC) and Symptom Identification (SI), as shown in Fig. 2. This structured design evaluates model performance by grouping these seven tasks into the three aforementioned hierarchical sub-domains:

- **Basic perception:** Provides basic classification of the presence of pathology (HDC), leaf condition (LC), and evaluates host crop species identification (CSI).
- **Pathology Understanding:** Focuses on fine-grained manifestations such as spots, chlorosis, and mosaic symptoms (SI), which help identify the pathogen category (PC).
- **Expert Diagnosis:** Target specific pathological condition (DC) and predict the scientific name of the disease on the leaf (SNC).

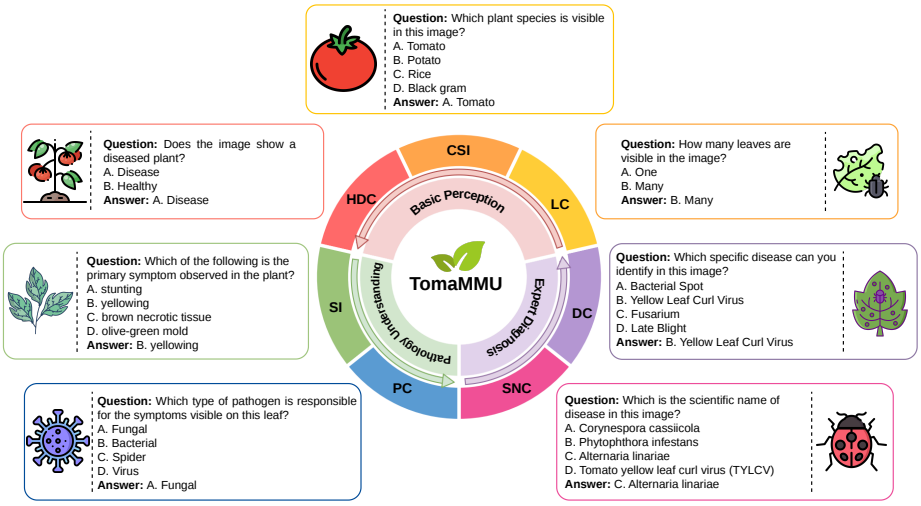


Fig. 2: Visualization of the question tasks distribution in VQA benchmark dataset. Each colored segment represents one of seven question types. The dataset includes HDC, SI, DC, LC, CSI, SNC and PC.

Stage 3: QA Generation For rigorous evaluation, TomaMMU adopts MCQs and OEQs. Using the dataset, a small, balanced subset, TomaBench, is selected for evaluation, while the remaining data is used for fine-tuning and training. All questions are designed to be visual-dependent and ensure that the selected conversations closely resemble real-life conversations between farmers and experts. To balance across different agricultural domains and knowledge types, we prioritize an even distribution of question types, as shown in Fig. 3c. To generate MCQs, depending on the knowledge type of the fact, we generate three wrong answers and form one correct answer, using the extracted fact as the ground truth for the

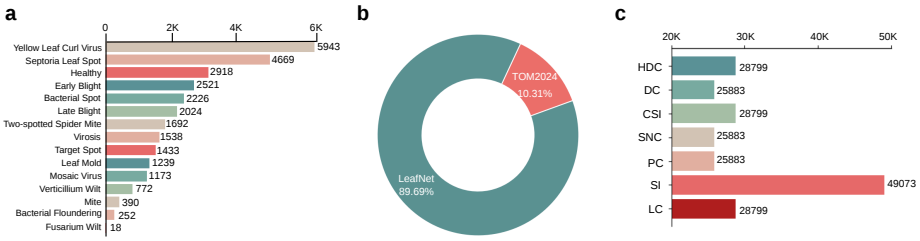


Fig. 3: Visualization of key components of TomaMMU. **a,** Distribution of number classes of disease in the TomaMMU. **b,** TomaMMU data sources. **c,** Number of VQA on different question types.

model to predict. With OEQs, the models are set up to offer short answers and uses Gemini 2.0 Flash [4] as a judge to validate generated answers and ground truth labels to specify semantic accuracy and correctness of VLMs.

3.3 Additional Properties of TomaMMU

Distribution and Coverage In Fig. 3, we show the distribution of the number of classes and knowledge types in TomaMMU. We have two key observations: (1) In the raw image distribution, natural factual events are heavily skewed toward a few major diseases, leaving other important conditions underrepresented, resulting in a severe imbalance across disease categories. To address this, we developed a structured VQA framework in which a single image is associated with up to seven distinct question types, ranging from basic perception to expert-level diagnosis. This strategy significantly enriches the density of supervision signals and balances the representation of agricultural knowledge. As demonstrated in Fig. 3c, these generated VQA conversations exhibit a markedly more uniform distribution across task categories, facilitating a more equitable and comprehensive assessment of VLMs capabilities. This is particularly important for our evaluation dataset, where we aim to propose a comprehensive standard that tests different aspects of the model without inductive bias.

Realistic Images Our TomaMMU is also different from photography-level images, laboratory datasets like PlantVillage, or textbook or web-document such as manually curated CROP [46] and CDDM [17]. TomaMMU is a conjunction of a large portion of data collected directly in the field (farm) with some data from a controlled environment (laboratory). Our benchmark also emphasizes training VLMs to align complex morphological patterns (such as sores, spots, colors) with in-depth textual descriptions.

4 Experimental Result

We conducted comprehensive evaluations to demonstrate the efficacy of domain-specific fine tuning in agricultural tasks and to establish a robust comparative

baseline in a variety of different VLMs. Primarily, we present refined results on TomaLLaVA, a multimodal generative framework that jointly leverages visual and textual information to perform accurate and interpretable tomato disease diagnosis. Furthermore, we benchmarked several state-of-the-art VLMs. Our evaluations focused primarily on zero-shot performance, using either publicly available APIs or author-provided checkpoints to reflect the inherent capabilities of each model without task-specific training. All experiments were conducted using NVIDIA RTX 4060 Ti GPUs, and more detailed information on model configuration, API choices, and implementation details is included in the supplementary documentation.

SOTA competing models. We assess a broad range of state-of-the-art VLMs to ensure robust comparisons. For open-sourced models, we prioritize architectures with comparable parameter numbers for a fair and meaningful comparison. We identify three categories of models: Proprietary VLMs, generative FMs, and CLIP-based models. *Proprietary VLMs* is Gemini 2.0 Flash [4], *General FMs.* We include eight VLMs (InternVL3 [42], Qwen2.5-3B [44], Qwen3-2B [43], Gemma4-E4B [18], SmolVLM2-2.2B [1], LLaVA1.5 [16] and LFM2-VL [2]). *CLIP-based models.* we utilize (CLIP [29], SigLIP2 [40]), and two *domain-specific VLMs* (BioCLIP [38] and SCOLD [28]), which are models refined from existing foundational models to optimize for specific tasks. All were evaluated without task-specific fine-tuning to assess out-of-the-box generalization.

4.1 Evaluation Metrics

To ensure robustness and fairness, we evaluate the models on MCQs and OEQs. **For MCQs**, we report Accuracy (Acc) and F1 score (F1) as the primary evaluation metric. We score the model's response by matching its predicted option to the ground truth.

For OEQs, we first utilize ROUGE-L [15] to comprehensively assess the lexical and structural similarity between generated responses and human-annotated data. Furthermore, we implement the LLM-as-judge methodology using the GPT score to grade answers semantically. The short-form responses correspond to the questions, which normally contain several words, and are evaluated directly against the correct answers. Specifically, our Large Language Models (LLM) judge follows a three-step process: (a) dividing both predicted and ground-truth responses into individual statements; (b) grading each of them as a short-form response; and (c) normalizing the grades according to the number of statements per question. This allows us to calculate the final harmonic mean of the semantic scores (1-5), which is designed to punish complete errors (score 1) severely. Detailed construction and specifications are provided in Section 4.4.

4.2 Zero-shot Evaluation Results

We present a comprehensive zero-shot comparison of various VLMs in Table 1. Evaluated specifically on the TomaBench for Generative FMs, CLIP-based models, Fine-tuned models, and Domain-specific VLMs.

Model	HDC		DC		CSI		SNC		PC		SI		LC		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Proprietary VLMs																
Gemini 2.0 Flash [4]	91.0	62.6	47.5	36.2	92.9	92.9	51.6	51.6	70.5	70.5	44.0	45.4	94.6	86.9	70.3	63.7
Generative VLMs																
SmolVLM2-2.2B [1]	36.3	23.0	33.3	31.3	36.3	33.4	42.5	42.6	40.1	37.7	36.4	35.5	92.9	56.3	45.4	37.1
Gemma4-E4B [18]	11.6	11.0	30.9	19.8	27.4	18.8	24.4	14.3	24.5	20.9	25.3	19.1	29.5	27.2	24.8	18.7
InternVL3-1B [42]	89.9	47.3	28.6	28.5	13.7	13.6	30.7	30.5	50.8	50.6	37.5	37.6	93.7	84.7	49.3	41.8
InternVL3-2B [42]	82.1	43.6	34.2	34.1	46.2	46.1	42.7	42.6	51.4	51.5	32.4	32.4	94.4	85.3	54.8	47.9
Qwen2.5-3B [44]	83.8	64.1	40.6	39.8	62.9	62.9	47.5	47.3	51.0	51.0	32.5	32.3	92.6	52.6	58.7	50.0
Qwen3-2B [43]	91.6	48.0	45.6	44.2	60.9	61.0	32.6	32.6	55.9	55.2	33.9	32.3	93.7	84.2	59.2	51.1
LLaVA-1.5-7B [16]	89.9	47.3	34.8	23.6	50.2	49.1	24.8	10.3	45.2	42.5	26.7	16.4	93.2	80.1	52.1	38.5
LFM2.5-VL-1.6B [2]	90.9	75.3	41.9	41.6	76.5	76.4	38.1	35.3	54.1	53.8	33.0	29.7	94.3	42.8	61.3	50.7
CLIP-based Models																
SigLIP2 [40]	76.1	63.7	25.6	25.2	30.1	30.1	27.8	27.8	11.2	11.1	28.6	28.6	93.1	81.1	41.8	38.2
CLIP-ViT-L-14 [29]	90.2	78.1	36.9	36.5	54.2	54.2	29.7	29.7	51.3	51.2	26.1	26.1	59.7	52.4	49.7	46.9
SCOLD [28]	55.1	48.7	74.6	74.2	**98.0**	**98.0**	49.3	49.3	47.4	47.4	52.1	52.1	44.9	39.9	60.2	58.5
BioCLIP-ViT-B-32 [38]	90.0	51.1	44.5	26.5	56.4	2.7	43.4	27.0	40.8	20.0	23.2	15.4	93.2	80.4	55.9	31.9
TomaLLaVA	**98.9**	**98.9**	**96.4**	**96.4**	-	-	**95.5**	**95.5**	**98.9**	**98.9**	**91.2**	**88.7**	**99.2**	**99.2**	**96.1**	**95.8**

Table 1: Performance of VLMs on our TomaBench MCQs. Our evaluation set poses great challenges to existing large VLMs. TomaLLaVA ignores the CSI task for non-image crops in TomaBench.

This presents significant challenges for existing models Beyond simple classification, diagnostic support systems must possess reasoning capabilities. TomaBench proved to be a very challenging benchmark for all evaluated models, and even the most advanced systems achieved average performance levels as shown in Table 2. These challenges are noteworthy, particularly for OEQs, where models must accurately recall knowledge without fine-tuning on question-answer pairs. These observations also highlight the need to improve the agricultural expertise of VLMs and to incorporate agricultural data into VLMs training, where our TomaMMU could be useful.

Performance Across Question Types On MCQs, We assessed model performance on closed-form QA multiple-choice questions, which are significantly easier than OEQs. This protocol tests the fit between visual features and complex agricultural linguistic concepts, measuring the model's ability to perform expert-level inference without task-specific training. The level of accuracy is also higher than the agricultural standards set by the OEQs in the responses. However, there is significant variation among models in which tasks are the most difficult. In particular, PC, SI, and SNC tasks struggle with agricultural diagnostics. This reflects the inherent difficulty of these tasks, which requires the model to process detailed morphological descriptions. In contrast, binary HDC and LC maintain a concise and uniform prompt structure, consistent with their high-level screening objective. Suggests that specialized agricultural knowledge and visual understanding capabilities are not uniformly distributed across model architectures.

	TomaBench-OEQs								
	HDC	DC	CSI	SNC	PC	SI	LC	Avg.	GPT
SmolVLM2-2.2B [1]	0.458	0.146	0.043	0.035	0.348	0.050	0.911	0.284	∼2.3
Gemma4-E4B [18]	0.694	0.191	0.017	0.005	0.323	0.073	0.707	0.287	∼2.3
InternVL3 1B [42]	0.816	0.016	0.001	0.001	0.381	0.142	0.840	0.314	∼2.1
InternVL3 2B [42]	0.874	0.255	0.055	0.060	0.357	0.065	0.905	0.367	∼2.2
Qwen2.5 3B [44]	0.790	0.234	0.034	0.048	0.267	0.202	0.796	0.339	∼2.4
Qwen3 2B [43]	0.899	0.182	0.341	0.099	0.071	0.203	0.869	0.380	∼2.7
LLaVA-OV [16]	0.914	0.023	0.111	0.000	0.005	0.071	0.906	0.290	∼2.6
LFM-VL 2.5 [2]	0.929	0.141	0.715	0.019	0.319	0.166	0.895	0.455	∼2.6

Table 2: Performance of VLMs on our TomaBench OEQs. Our evaluation set poses great challenges to existing large VLMs.

4.3 Fine-tuning Result

Although agricultural VLMs have made progress, they still fail to exploit the specific characteristics of individual diseases or species. As a result, these models often lack the detailed domain knowledge required to distinguish apparently similar tomato diseases and to support reasoning-based diagnosis across diverse real-world agricultural conditions, and tomato leaves disease data is the initial foundation. Notably, even highly advanced foundation models such as Gemini 2.0 Flash [4] only achieved 70.30% on the overall dataset; therefore, the need for a model fine-tuned specifically on this dataset was paramount. This performance gap illustrates the requirement to train a domain model for TomaMMU. This comparative evaluation demonstrates TomaLLaVA's superior ability to generate accurate, contextually rich disease diagnoses and symptom explanations beyond what standard classification-based approaches can provide. In Fig. 4 summarizes the performance of all models and top models across seven question types. Overall, TomaLLaVA achieves the highest performance across all tasks, with an average Acc of 96.09% and an F1 of 0.958, substantially outperforming all competing models. As shown in Table 1. Compared with the best VLMs (LFM2.5, 61.26% Acc, 0.507 F1), TomaLLaVA gains +34.8% Acc and +0.451 F1, even in Fine-tuned VLMs (AgriCLIP, 93.73% Acc, 0.948 F1), TomaLLaVA gains +2.4% Acc and +0.01 F1. Overall, the excellent TomaLLaVA results cover all models and deliver near-perfect results for perception tasks and strong fine-grained disease reasoning, confirming its ability to emulate expert-level diagnostic reasoning through multimodal integration.

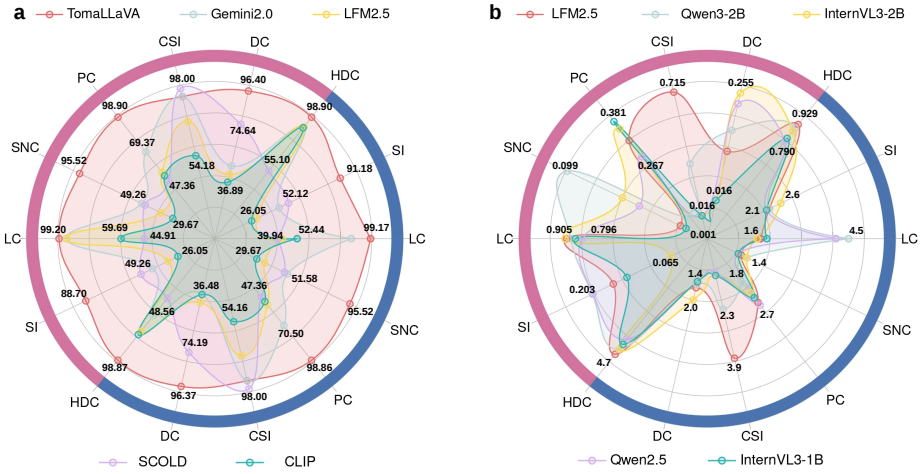


Fig. 4: Summary of model results on MCQs and OEQs. **a,** Performance of various foundation models across seven tomato disease understanding tasks (**pink** indicates Acc, and **blue** denotes F1). Performance of Top-5 SOTA models on MCQs, showing that TomaLLaVA maintains the highest overall performance. b, Comparison of the Top-5 state-of-the-art models on OEQs (**pink** indicates the average of ROUGE-L, and **blue** denotes the GPT score.

4.4 Error Analysis for OEQs

To achieve an in-depth understanding of the internal mechanisms and limitations of VLMs when answering OEQs, we conducted automated error analysis and qualitative evaluation. Unlike some previous studies that relied entirely on manual scoring, we applied an LLM-as-a-judge method, using the GPT score to evaluate and score model predictions based on the reference answer (Ground Truth) on a scale of 1 to 5. Our prompt for Gemini 2.0 Flash is presented in Table S2.

Table 2 showed that all models encountered limitations and achieved only scores $\approx$ 2.5. Therefore, we classified and clarified the nature of the errors corresponding to these three score scenarios.

Perceptual Error & Refusal to Answer (Score 1) This delegation for dead failure in the process question type: the VLMs lacks knowledge or visual grounding, giving completely wrong, irrelevant, or hallucinatory answers. For example, mistaking "*Tomato*" for "*Eucalyptus*" or "*Virus*" for "*Bacteria*". Moreover, instead of an analytical image, Gemma4-E4B [18] repeated allegations, stereotypes, and negative responses such as "*No plant is visible*" for SI or "*No visible disease*" for DC, even though the image clearly shows signs of the disease. The system failure rendered the predictions completely worthless.

Core Miss & Major Inaccuracy (Score 2) The answer is partially relevant but severely inaccurate or misses the core idea. In these cases, the model may identify relevant visual features and strengthen them, but still fail to reach the correct answer (e.g., describing "*Brown Spot*" instead of "*Early Leaf Blight*").

In other cases, the model only identified secondary symptoms "*wilting*" while overlooking the dangerous primary symptoms "*death*" and "*stunting*".

Granularity & Detail Omission (Score 3) The answer is generally correct but lacks specific details or contains minor factual errors. This category highlights limitations in deep agricultural knowledge. For instance, recognizing discoloration "*yellowing*" but overlooking accompanying structural deformation "*curling*", or correctly identifying the type of damage "*leaf spots*" but lacking detail about the color "*reddish-brown*".

4.5 Training result

Experiment Setting We conducted training TomaLLaVA, which uses SCOLD [28] as the image encoder and LFM2.5-1.2B-Instruct [2] as the base LLM. Our training lasts 20 epochs using Low-Rank Adaptation (LoRA) [11], a learning rate of 1e-4, a weight decay of 0.01, and a batch size of 32. The fine-tuning process takes on NVIDIA RTX 4060 Ti GPUs.

Fine-tuning Performance. Fig. 5 shows that the fine-tuning with our knowledge base generates a significant performance boost, improving the capability of understanding the images and correctly responding with agriculture knowledge, leading to an average improvement of up to 2.36% on MCQs. This performance boost demonstrates the effectiveness of our large-scale knowledge base in improving VLMs and the demand to collect agricultural-related data for future VLMs. However, for questions on pathology understanding and expert diagnosis, the level of improvement is smaller, indicating the complexity of multi-statement questions and underscoring the need for more data collection and better model training beyond simple fine-tuning.

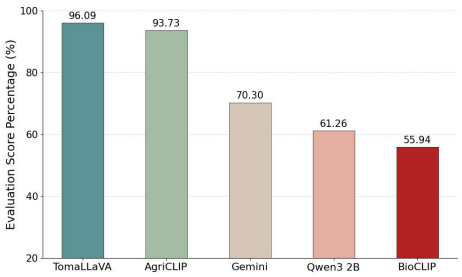


Fig. 5: Average evaluation scores on MCQs. Fine-tuning TomaLLaVA significantly enhances its understanding of agricultural knowledge, demonstrating the effectiveness of our development set compared with four state-of-the-art models.

Failure Exemplars. Fig. 6 indicated that the most persistent bottlenecks lie in agronomic reasoning, which requires interpretation of biological and agronomic context, such as symptoms, pathogens, and disease progression. These limitation stem from general datasets lacking specialized knowledge, diversity, and realism, making the models susceptible to illusions and mispredictions. For instance, when asked to identify a specific disease from an image of a leaf with characteristic necrotic spots, generals VLMs lacked of agronomic refinement gave incorrect diagnoses such as SCOLD [28] predicted "*Verticillium Wilt*" and Gemini 2.0 Flash [4] predicted "*Leaf Mold*". In contrast, TomaLLaVA accurately applied specialized knowledge to correctly identify "*Early Blight*". Additionally, for simple tasks like LC, foundation models such as SCOLD and Genini 2.0 Flash still

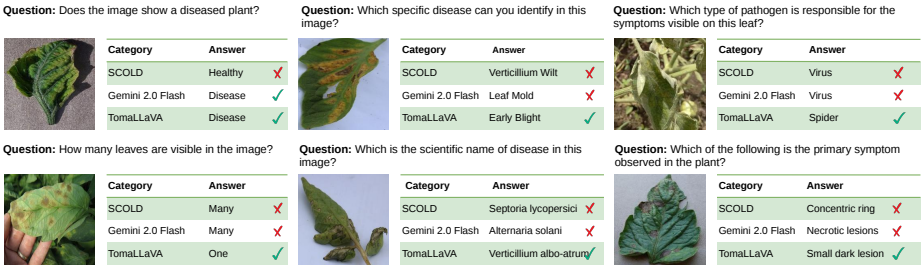


Fig. 6: The effectiveness of TomaLLaVA fine-tuning on OEQ examples. After simple fine-tuning, TomaLLaVA can accurately identify issues that SCOLD [28] and Gemini 2.0 Flash [4] failed to recognize in zero-shot scenarios.

focus more on the background than on the main object, as their predictions are ”Many".

5 Conclusion

In this study, we introduce TomaMMU, a large-scale dataset and benchmark for tomato disease VLMs in agriculture, a field that requires high levels of expert knowledge and accurate interpretation of visuals. Our dataset comprises over 213,119 human-annotated image metadata pairs. Applying a three-stage curation framework, we construct a high-quality TomaMMU evaluation set and demonstrate its crucial role in improving performance of multimodal models. To support model development, we propose TomaBench set, including 42,626 question-answer pairs to evaluate multimodal ability. Our contributions encompass not only the dataset but also a structured fine-tuning strategy and a human-annotated gold standard split to foster future research in agricultural technology and improve the capabilities of AI systems with complex reasoning requirements for tomato leaves disease, as well as a platform to build and deploy to other datasets in agriculture.

Acknowledgment

Khang Nguyen Quoc was supported by the Hyundai Motor Chung Mong-Koo Foundation Global Scholarship (GSS-25-02120).

References

1. Allal, L.B., Lozhkov, A., Bakouch, E., Blázquez, G.M., Penedo, G., Tunstall, L., Marafioti, A., Kydlíček, H., Lajarín, A.P., Srivastav, V., Lochner, J., Fahlgren, C., Nguyen, X.S., Fourrier, C., Burtenshaw, B., Larcher, H., Zhao, H., Zakka, C., Morlon, M., Raffel, C., von Werra, L., Wolf, T.: Smollm2: When smol goes big

– data-centric training of a small language model (2025). https://doi.org/10.48550/ARXIV.2502.02737, https://arxiv.org/abs/2502.02737 9, 10, 11
2. Amini, A., Banaszak, A., Benoit, H., Böök, A., Dakhran, T., Duong, S., Eng, A., Fernandes, F., Härkönen, M., Harrington, A., Hasani, R., Karwa, S., Khrustalev, Y., et al.: Lfm2 technical report (2025). https://doi.org/10.48550/ARXIV.2511.23404, https://arxiv.org/abs/2511.23404 9, 10, 11, 13
3. Appiah, O., Hackman, K.O., Diallo, B.A.A., Ogunjobi, K.O., Ouedraogo, V., Bebe, M., SON, D.: Tom2024 (2024). https://doi.org/10.17632/3D4YG89RTR.1, https://data.mendeley.com/datasets/3d4yg89rtr/1 6
4. Balestri, R.: Gender and content bias in large language models: a case study on google gemini 2.0 flash experimental. Frontiers in Artificial Intelligence **8** (Mar 2025). https://doi.org/10.3389/frai.2025.1558696, http://dx.doi.org/10.3389/frai.2025.1558696 8, 9, 10, 11, 13, 14
5. Czosnek, H., Laterrot, H.: A worldwide survey of tomato yellow leaf curl viruses. Archives of Virology **142**(7), 1391–1406 (1997). https://doi.org/10.1007/s007050050168, http://dx.doi.org/10.1007/s007050050168 1
6. George, R., Thuseethan, S., Ragel, R.G., Mahendrakumaran, K., Nimishan, S., Wimalasooriya, C., Alazab, M.: Past, present and future of deep plant leaf disease recognition: A survey. Computers and Electronics in Agriculture **234**, 110128 (2025). https://doi.org/10.1016/j.compag.2025.110128, http://dx.doi.org/10.1016/j.compag.2025.110128 2
7. GLASSCOCK, W.H., WARE, W.M.: Alternaria solani on tomato. Nature **154**(3916), 642–642 (Nov 1944). https://doi.org/10.1038/154642a0, http://dx.doi.org/10.1038/154642a0 1
8. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering (2016). https://doi.org/10.48550/ARXIV.1612.00837, https://arxiv.org/abs/1612.00837 3
9. Grbić, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouzé, P., Grbić, V., Osborne, E.J., et al.: The genome of tetranychus urticae reveals herbivorous pest adaptations. Nature **479**(7374), 487–492 (Nov 2011). https://doi.org/10.1038/nature10640, http://dx.doi.org/10.1038/nature10640 1
10. Gupta, A., Kamalachandran, D., Longchar, B., Senthil-Kumar, M.: Impact of soil moisture regimes on wilt disease in tomatoes: Current understanding p. 73–82 (2018). https://doi.org/10.1016/b978-0-12-813066-7.00005-x, http://dx.doi.org/10.1016/B978-0-12-813066-7.00005-X 1
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021). https://doi.org/10.48550/ARXIV.2106.09685, https://arxiv.org/abs/2106.09685 13
12. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering (2019). https://doi.org/10.48550/ARXIV.1902.09506, https://arxiv.org/abs/1902.09506 3
13. Jafar, A., Bibi, N., Naqvi, R.A., Sadeghi-Niaraki, A., Jeong, D.: Revolutionizing agriculture with artificial intelligence: plant disease detection methods, applications, and their limitations. Frontiers in Plant Science **15** (Mar 2024). https://doi.org/10.3389/fpls.2024.1356260, http://dx.doi.org/10.3389/fpls.2024.1356260 2
14. Khan, S.U., Alsuhaibani, A., Alabduljabbar, A., Almarshad, F., Altherwy, Y.N., Akram, T.: A review on automated plant disease detection: motivation, limitations, challenges, and recent advancements for future research. Journal of King Saud

University Computer and Information Sciences **37**(3) (May 2025). https://doi.org/10.1007/s44443-025-00040-3, http://dx.doi.org/10.1007/s44443-025-00040-3 2
15. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/ 9
16. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023). https://doi.org/10.48550/ARXIV.2304.08485, https://arxiv.org/abs/2304.08485 9, 10, 11
17. Liu, X., Liu, Z., Hu, H., Chen, Z., Wang, K., Wang, K., Lian, S.: A Multimodal Benchmark Dataset and Model for Crop Disease Diagnosis, p. 157–170. Springer Nature Switzerland (Oct 2024). https://doi.org/10.1007/978-3-031-73016-0_10, http://dx.doi.org/10.1007/978-3-031-73016-0_10 4, 8
18. Manik, M.M.H., Wang, G.: Unified deployment-aware evaluation of open reasoning language models (2026). https://doi.org/10.48550/ARXIV.2604.07035, https://arxiv.org/abs/2604.07035 9, 10, 11, 12
19. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Frontiers in Plant Science **7** (2016). https://doi.org/10.3389/fpls.2016.01419, http://dx.doi.org/10.3389/fpls.2016.01419 2
20. Moupojou, E., Tagne, A., Retraint, F., Tadonkemwa, A., Wilfried, D., Tapamo, H., Nkenlifack, M.: Fieldplant: A dataset of field plant images for plant disease detection and classification with deep learning. IEEE Access **11**, 35398–35410 (2023). https://doi.org/10.1109/access.2023.3263042, http://dx.doi.org/10.1109/ACCESS.2023.3263042 2
21. Naik, A., Chaudhary, A.: Tomato Mosaic Virus, p. 329–347. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-031-81884-4_21, http://dx.doi.org/10.1007/978-3-031-81884-4_21 1
22. Nawaz, U., Awais, M., Gani, H., Naseer, M., Khan, F., Khan, S., Anwer, R.M.: Agriclip: Adapting clip for agriculture and livestock via domain-specialized cross-model alignment (2024). https://doi.org/10.48550/ARXIV.2410.01407, https://arxiv.org/abs/2410.01407 5
23. Nowicki, M., Kozik, E.U., Foolad, M.R.: Late blight of tomato (Oct 2013). https://doi.org/10.1002/9781118728475.ch13, http://dx.doi.org/10.1002/9781118728475.ch13 1
24. Osdaghi, E.: Xanthomonas vesicatoria (bacterial spot of tomato and pepper) (Dec 2020). https://doi.org/10.1079/cabicompendium.56981, http://dx.doi.org/10.1079/cabicompendium.56981 1
25. Pandey, A., Paudel, R., Adhikari, T.B., Panthee, D.R., Louws, F.J.: Septoria leaf spot of tomatoes: Historical insights, present challenges, and future prospects. Horticulturae **10**(12), 1299 (Dec 2024). https://doi.org/10.3390/horticulturae10121299, http://dx.doi.org/10.3390/horticulturae10121299 1
26. Quach, L.D., Quoc Khang, N., Thai-Nghe, N., Nguyen, C.N.: Xai-bo: an architecture using grad-cam technique to evaluate bayesian optimization algorithms on deep learning models. Journal of Information and Telecommunication **9**(3), 335–356 (Jan 2025). https://doi.org/10.1080/24751839.2024.2447191, http://dx.doi.org/10.1080/24751839.2024.2447191 2
27. Quoc, K.N., Dao, P.D., Quach, L.D.: Leafnet: A large-scale dataset and comprehensive benchmark for foundational vision-language understanding of plant diseases (2026). https://doi.org/10.48550/ARXIV.2602.13662, https://arxiv.org/abs/2602.13662 3, 4, 6

28. Quoc, K.N., Thu, L.L.T., Quach, L.D.: A vision-language foundation model for leaf disease identification (2025). https://doi.org/10.48550/ARXIV.2505.07019, https://arxiv.org/abs/2505.07019 5, 9, 10, 13, 14
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021). https://doi.org/10.48550/ARXIV.2103.00020, https://arxiv.org/abs/2103.00020 9, 10
30. Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., Bhattacharyya, P.: Scienceqa: a novel resource for question answering on scholarly articles. International Journal on Digital Libraries **23**(3), 289–301 (2022). https://doi.org/10.1007/s00799-022-00329-y, http://dx.doi.org/10.1007/s00799-022-00329-y 3
31. Sajitha, P., Andrushia, A.D., Anand, N., Naser, M.: A review on machine learning and deep learning image-based plant disease classification for industrial farming systems. Journal of Industrial Information Integration **38**, 100572 (Mar 2024). https://doi.org/10.1016/j.jii.2024.100572, http://dx.doi.org/10.1016/j.jii.2024.100572 2
32. Sakib, S.N., Haque, N., Hossain, M.Z., Arman, S.E.: Plantvillagevqa: A visual question answering dataset for benchmarking vision-language models in plant science (2025). https://doi.org/10.48550/ARXIV.2508.17117, https://arxiv.org/abs/2508.17117 6
33. Schlub, R., Smith, L., Datnoff, L., Pernezny, K.: An overview of target spot of tomato caused by corynespora cassiicola. Acta Horticulturae (808), 25–28 (Jan 2009). https://doi.org/10.17660/actahortic.2009.808.1, http://dx.doi.org/10.17660/ActaHortic.2009.808.1 1
34. Senthil Pandi, S., Senthilselvi, A., Gitanjali, J., ArivuSelvan, K., Gopal, J., Vellingiri, J.: Rice plant disease classification using dilated convolutional neural network with global average pooling. Ecological Modelling **474**, 110166 (Dec 2022). https://doi.org/10.1016/j.ecolmodel.2022.110166, http://dx.doi.org/10.1016/j.ecolmodel.2022.110166 2
35. Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., Batra, N.: Plantdoc: A dataset for visual plant disease detection. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. p. 249–253. CoDS COMAD 2020, ACM (Jan 2020). https://doi.org/10.1145/3371158.3371196, http://dx.doi.org/10.1145/3371158.3371196 2
36. SINHA, A.K., WOOD, R.K.S.: Studies on the nature of resistance in tomato plants to verticillium albo-atrum. Annals of Applied Biology **62**(2), 319–327 (Oct 1968). https://doi.org/10.1111/j.1744-7348.1968.tb02827.x, http://dx.doi.org/10.1111/j.1744-7348.1968.tb02827.x 1
37. Srinivas, C., Nirmala Devi, D., Narasimha Murthy, K., Mohan, C.D., Lakshmeesha, T., Singh, B., Kalagatur, N.K., Niranjana, S., Hashem, A., Alqarawi, A.A., Tabassum, B., Abd_Allah, E.F., Chandra Nayaka, S., Srivastava, R.K.: Fusarium oxysporum f. sp. lycopersici causal agent of vascular wilt disease of tomato: Biology to diversity– a review. Saudi Journal of Biological Sciences **26**(7), 1315–1324 (Nov 2019). https://doi.org/10.1016/j.sjbs.2019.06.002, http://dx.doi.org/10.1016/j.sjbs.2019.06.002 1
38. Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carlyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T., Chao, W.L., Su, Y.: Bioclip: A vision foundation model for the tree of life (2023). https://doi.org/10.48550/ARXIV.2311.18803, https://arxiv.org/abs/2311.18803 5, 9, 10

39. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms (2024). https://doi.org/10.48550/ARXIV.2401.06209, https://arxiv.org/abs/2401.06209 4
40. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features (2025). https://doi.org/10.48550/ARXIV.2502.14786, https://arxiv.org/abs/2502.14786 9, 10
41. 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 (Nov 2023). https://doi.org/10.1038/s43016-023-00867-x, http://dx.doi.org/10.1038/s43016-023-00867-x 3
42. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models (2025). https://doi.org/10.48550/ARXIV.2504.10479, https://arxiv.org/abs/2504.10479 9, 10, 11
43. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., et al.: Qwen3 technical report (2025). https://doi.org/10.48550/ARXIV.2505.09388, https://arxiv.org/abs/2505.09388 9, 10, 11
44. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., et al.: Qwen2.5 technical report (2024). https://doi.org/10.48550/ARXIV.2412.15115, https://arxiv.org/abs/2412.15115 9, 10, 11
45. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi (2023). https://doi.org/10.48550/ARXIV.2311.16502, https://arxiv.org/abs/2311.16502 4
46. Zhang, H., Sun, J., Chen, R., Liu, W., Yuan, Z., Zheng, X., Wang, Z., Yang, Z., Yan, H., Zhong, H.S., Wang, X., Ouyang, W., Yang, F., Dong, N.: Empowering and assessing the utility of large language models in crop science. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024), https://openreview.net/forum?id=hMj6jZ6JWU 8
47. Zhao, T., Pei, T., Jiang, J., Yang, H., Zhang, H., Li, J., Xu, X.: Understanding the mechanisms of resistance to tomato leaf mold: A review. Horticultural Plant Journal **8**(6), 667–675 (Nov 2022). https://doi.org/10.1016/j.hpj.2022.04.008, http://dx.doi.org/10.1016/j.hpj.2022.04.008 1
48. Zhou, Y., Ryo, M.: AgriBench: A Hierarchical Agriculture Benchmark for Multimodal Large Language Models, p. 207–223. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-031-91835-3_14, http://dx.doi.org/10.1007/978-3-031-91835-3_14 4
49. Zhu, H., Qin, S., Su, M., Lin, C., Li, A., Gao, J.: Harnessing large vision and language models in agriculture: a review. Frontiers in Plant Science **16** (2025). https://doi.org/10.3389/fpls.2025.1579355, http://dx.doi.org/10.3389/fpls.2025.1579355 2

TomaMMU: A Comprehensive Multimodal Understanding Benchmark for Tomato Leaf Diseases

– Supplementary Material –

A Reference Papers for Agronomic Reasoning

Table S1: Reference papers utilized for the agronomic reasoning of tomato diseases.

Disease Name	Reference Paper
Bacterial Spot	Xanthomonas vesicatoria (bacterial spot of tomato and pepper) [24]
Bacterial Wilt	Impact of Soil Moisture Regimes on Wilt Disease in Tomatoes: Current Understanding [10]
Early Blight	Alternaria Solani on Tomato [7]
Late Blight	Late Blight of Tomato [23]
Fusarium Wilt	Fusarium oxysporum f. sp. lycopersici causal agent of vascular wilt disease of tomato: Biology to diversity– A review [37]
Verticillium Wilt	Studies on the nature of resistance in tomato plants to Verticillium albo-atrum [36]
Leaf Mold	Understanding the mechanisms of resistance to tomato leaf mold: A review [47]
Septoria Leaf Spot	Septoria Leaf Spot of Tomatoes: Historical Insights, Present Challenges, and Future Prospects [25]
Target Spot	An overview of target spot of tomato caused by corynespora cassiicola [33]
Yellow Leaf Curl Virus	A worldwide survey of tomato yellow leaf curl viruses [5]
Mosaic Virus	Tomato mosaic virus [21]
Spider Mites	The genome of Tetranychus urticae reveals herbivorous pest adaptations [9]

B Prompt Template for GPT Evaluation

To evaluate the performance of our candidate models and compute the GPT score, we utilize the structured evaluation prompt presented in Table S2.

Table S2: Prompt template utilized for the automated evaluation of prediction accuracy.

System Prompt & Role

You are an expert multimodal AI evaluator. Your task is to evaluate the accuracy of a candidate answer to a visual question, based on the provided ground truth answer set.

[Input Data]

- **Ground Truth Reference Answers:** {ground_truth_answers}
- **Prediction Answer:** {prediction_answer}

[Evaluation Criteria]

Evaluate the candidate answer based on the following scale (1 to 5):

1: The answer is completely incorrect, hallucinatory, or irrelevant to the question.

2: The answer is partially relevant but contains major inaccuracies or misses the core question.

3: The answer is generally correct but lacks specific details, or contains minor factual errors.

4: The answer is correct, directly addresses the question, and is semantically equivalent to the ground truth.

5: The answer is exceptionally accurate, clear, and demonstrates a strong understanding of the visual and textual context.

[Output Format]

Provide your evaluation and give me the average score of the model.