Benchmarking Vision-Language Models for Microscopic Plant Image Understanding

Tianqi Wei[1], Xin Yu[2], Zhi Chen[3], Scott Chapman[4], and Zi Huang[1]

[1] The University of Queensland, Brisbane, Australia
{tianqi.wei,scott.chapman,helen.huang}@uq.edu.au
[2] Adelaide University, Adelaide, Australia
xin.yu@adelaide.edu.au
[3] University of Southern Queensland, Toowoomba, Australia
Zhi.Chen@unisq.edu.au

Abstract. Microscopic imaging provides essential visual evidence for studying plant biology and pathology at the cellular and subcellular levels. However, existing benchmarks for vision-language models primarily focus on macroscopic plant imagery, while the microscopic domain remains underexplored. To address this gap, we present PlantMicro, a comprehensive benchmark for evaluating vision–language models (VLMs) in microscopic plant imagery. PlantMicro integrates more than 5,000 images collected from diverse hosts, biological domains, and imaging modalities. Building on this diversity, we design a set of complementary tasks that capture different aspects of microscopic image understanding. To support these tasks, we construct over 9,000 VQA pairs that systematically evaluate the capabilities of VLMs. Experiments on PlantMicro show that current VLMs struggle with fine-grained recognition and biologically grounded reasoning. For example, GPT-5 achieves 34.93% accuracy on the pathogen classification task, which is only modestly above a 24.95% random guessing baseline. The results highlight a significant gap in the ability of current VLMs to comprehend microscopic plant images. PlantMicro provides a standardized foundation for advancing VLMs toward reliable and comprehensive microscopy-level plant understanding. PlantMicro is available at https://github.com/tqwei05/PlantMicro.

1 Introduction

Microscopy opens a window onto cellular and subcellular structure, revealing tissue organization, cell morphology, and fine biological details that are invisible at the field scale. For plant science, these features underpin explanations of development [36, 42, 44], stress adaptation [10, 18, 21], and host–pathogen interaction [33, 48]. Yet most evaluations of Vision–Language Models (VLMs) [4, 16, 27, 45, 52] in plants focus on macroscopic scenes [3, 13, 30, 49, 55]. As a result, microscopic visual understanding and its associated domain-specific terminology remain largely unexplored in VLMs.

Plants are fundamental to life on Earth. They drive primary productivity, release oxygen, and regulate global carbon cycling [11]. These roles make plant

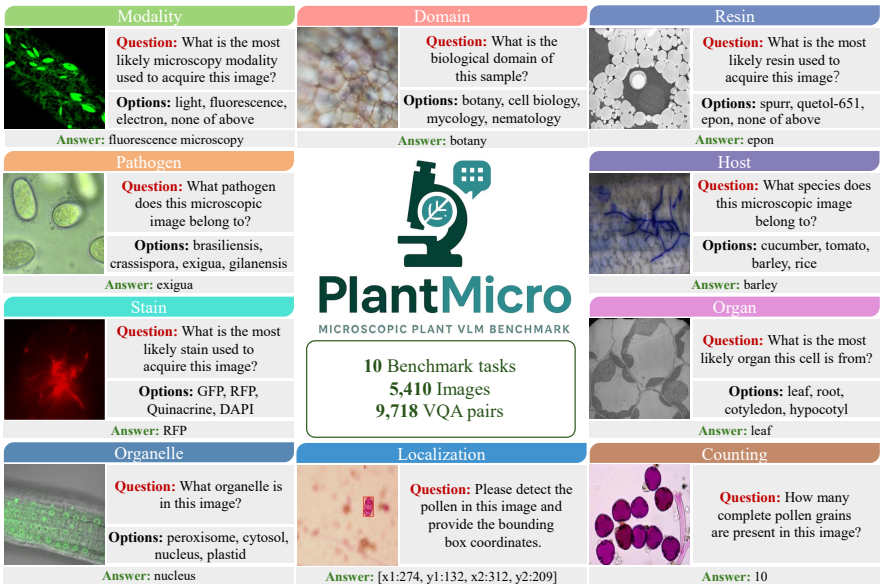


Fig. 1: VQA samples from different tasks in PlantMicro. PlantMicro consists of 5,410 microscopy images and 9,718 VQA pairs designed for microscopic understanding tasks.

science central to agriculture, ecology, and modern biology, with direct implications for food security and environmental sustainability [43]. With advances in computer vision, plant research has progressed rapidly across tasks such as species classification [28, 62, 66], disease detection [2, 58, 61], and growth-stage recognition [24, 38].

VLMs have emerged as general-purpose learners that link visual content with natural language processing [4, 16, 27, 45, 52]. Through prompt-based interaction they support open-vocabulary recognition [45], zero-shot classification [23, 63], and few-shot learning [1, 12]. Compared with conventional supervised models, VLMs provide an interpretable framework that connects perception with reasoning through language and can perform various tasks without task-specific training. In plant science, existing VLM benchmarks, *e.g.*, CDDM [30] and AgroBench [49], emphasize field-level imagery and operations such as weed identification, disease recognition, and management assistance, but they largely probe external appearance rather than the internal structures that microscopy exposes.

To bridge this gap, we introduce PlantMicro, a comprehensive benchmark for systematically evaluating VLMs on microscopic plant image understanding. PlantMicro spans multiple biological domains, including mycology, nematology, botany, cell biology, and incorporates various modalities, such as light, fluorescence, and electron microscopy. We define a diverse set of benchmark tasks spanning attribute prediction, biological entity identification, localization, and counting. Specifically, image-level attribute tasks include domain, modality, stain, and resin prediction. Biological entity identification tasks cover host, pathogen, or-

Table 1: Comparison of image benchmarks for plant science.

Benchmarks	#Host	#Images	#VQA Pairs	Expert	Micro.	Imaging Modality		
						Light	Fluor.	Electron
Tomato Spore [19]	1	400	–	✓	✓	✓	–	–
PODB [34]	13	459	–	✓	✓	–	✓	–
Hydrous Plants [51]	13	218	–	✓	✓	–	–	✓
CDDM [30]	15	137k	1M	–	–	–	–	–
AgroBench [49]	311	3,745	4,342	✓	–	–	–	–
Agri-CM^3 [55]	11	3,939	15,901	✓	–	–	–	–
PlantMicro (ours)	25	5,410	9,718	✓	✓	✓	✓	✓

ganelle, and organ recognition. In addition, we introduce localization to assess spatial grounding ability and target counting to evaluate model robustness in microscopic scenarios. PlantMicro adopts a visual question answering format for evaluation. All questions and annotations are derived from expert-level metadata in the source microscopy datasets and further refined through manual curation to ensure accurate and consistent labeling. Compared with existing benchmarks that either lack microscopy or do not support VLM benchmarking, PlantMicro allows fine-grained biological interpretation through expert-curated VQA pairs, as summarized in Table 1. Our main contributions can be summarized as:

- We present PlantMicro, a unified benchmark comprising 5,410 plant microscopy images and 9,718 visual question answering (VQA) pairs.
- We establish a suite of benchmark tasks derived from the standardized metadata, which provide a systematic evaluation for VLMs across diverse aspects of microscopic plant image understanding.
- We benchmark a wide range of VLMs on PlantMicro and provide a comprehensive analysis across diverse microscopic understanding tasks, revealing that VLMs are capable of generic perceptual classification tasks but struggle with fine-grained biological recognition, counting and localization tasks.

2 Related Work

Computer Vision in Plant Science. Computer vision has been widely applied in plant science for diverse visual perception and analysis tasks. Early studies focused on image classification, recognizing plant species or diseases [29, 54, 59]. Representative classification datasets like PlantVillage [37] provide clean, laboratory-collected images for disease identification, while PlantDoc [50] and PlantWild [58] offer in-the-wild samples with complex backgrounds and varying lighting conditions. Subsequent works extended to object detection, where deep models such as Faster R-CNN [47] and YOLO [46] enable localization and quantification of objects such as fruits, weeds, and insects in complex field environments [6, 25, 32, 53]. Segmentation methods further advance fine-grained understanding by predicting pixel-level regions for detailed analysis. These methods have been applied to diverse tasks, including leaf and canopy analysis, spike segmentation in cereals, and disease lesion delineation [14, 56, 60, 64]. Although traditional methods have advanced research in plant imagery, they still rely on

dedicated datasets, annotations, and model architectures, which limit scalability and generalization under different tasks.

Field-Level Plant Benchmarks for VLMs. Conventional vision models [2, 28, 61, 66] are task-specific and depend heavily on dedicated annotations, which limit their scalability and generalization. Vision–language models (VLMs) [4, 7, 39, 45] are trained on massive-scale multimodal data to align visual and textual information, enabling open-vocabulary recognition and reasoning. Inspired by their success in general domains, recent benchmarks have begun extending VLM evaluation to field-level plant imagery. To diagnose plant diseases, CDDM [30] constructed a large-scale multimodal benchmark with over 130k plant images and one million question–answer pairs. AgEval [3] covers twelve plant stress tasks under zero and few-shot settings, indicating that model performance depends strongly on how prompts are designed and whether the provided examples match the target domain. AgroBench [49] offers a large-scale evaluation across seven agricultural themes and finds that current VLMs still struggle with fine-grained recognition tasks such as weed and disease identification. Agri-CM^3 [55] proposes a hierarchical evaluation framework that assesses plant understanding across progressively complex levels of perception, reasoning, and knowledge application. MIRAGE [9] focuses on expert-guided multimodal conversations, where models need to interpret ambiguous inputs and produce practical responses. However, existing VLM benchmarks focus mostly on macroscopic plant imagery, leaving microscopic plant understanding largely underexplored.

Microscopic Benchmarks for VLMs. Recent work has begun extending VLMs into microscopic and biomedical domains. BiomedCLIP [65], pretrained on 15 million biomedical image–text pairs, provides a strong foundation for fine-grained tasks such as pneumonia detection. PMC-CLIP [26] leverages large-scale figure–caption data from biomedical literature and achieves strong performance in medical visual question answering. PLIP [15], trained on pathology images and text from Medical Twitter, learns diverse pathological patterns and diagnostic expressions from real-world data. Benchmarks such as MicroBench [31] have also been introduced to evaluate VLMs across microscopy modalities and biomedical tasks, covering light, fluorescence, and electron imaging for classification, captioning, and question answering. Despite these advances, most existing efforts primarily focus on human or animal biomedical imagery. In contrast, plant microscopy, with its distinct cellular organization and biological processes, remains an underexplored area. To fill this gap, we develop a unified framework for evaluating VLMs in microscopic plant understanding, bridging visual perception with plant biological reasoning.

3 The PlantMicro Benchmark

PlantMicro is a comprehensive benchmark for evaluating vision–language models (VLMs) on microscopic plant image understanding. We illustrate the construction process of our PlantMicro in Figure 2. It aggregates diverse microscopy images collected from public datasets in peer-reviewed publications and institutional repositories. The scope includes plant pathology, botany, cell biology, and

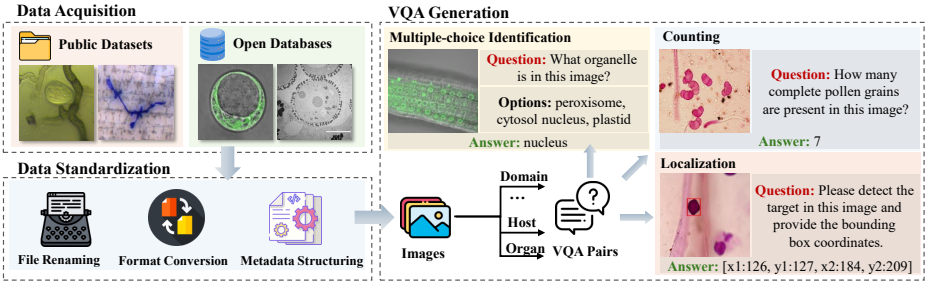
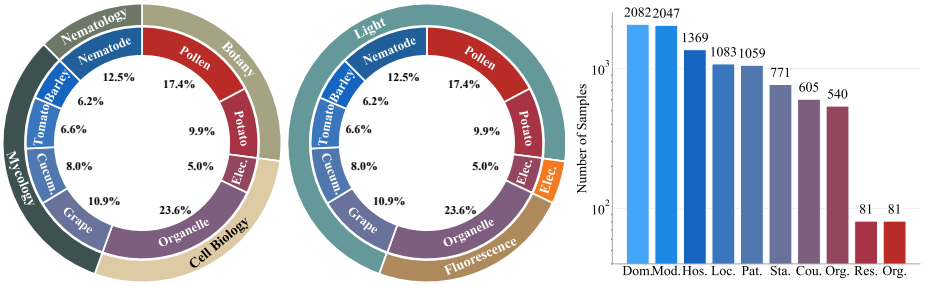


Fig. 2: Construction pipeline of the PlantMicro benchmark. The process includes data acquisition from public data sources, followed by standardized curation and metadata verification. Based on the validated attributes, we generate structured VQA pairs for multiple-choice identification, object localization, and target counting tasks.



(a) Domain distribution **(b)** Modality distribution **(c)** VQA task statistics

Fig. 3: Detailed statistics of **PlantMicro**. (a) Domain distribution. (b) Modality distribution. (c) The number of QA pairs of each benchmark task.

nematology. In addition, it covers a broad range of biological contexts and imaging modalities. We carefully curated, standardized, and annotated all images with unified metadata. Thus, the benchmark allows consistent evaluation across a wide range of tasks. In total, PlantMicro contains 9,718 VQA pairs derived from 5,410 microscopy images. Figure 3a and 3b illustrate the distribution of PlantMicro across biological domains and microscopy modalities, respectively. Figure 3c reports the number of VQA pairs for each benchmark task.

3.1 Data Review and Acquisition

We conduct a systematic search on peer-reviewed publications [5, 8, 17, 19, 33, 67, 68], and open data repositories [34, 35] to collect microscopy datasets of plant tissues and diseases. To ensure data diversity, we include multiple biological domains, such as mycology, botany, and nematology, as well as imaging modalities, including light, fluorescence, and electron microscopy. We prioritize datasets that are openly available for research use. For those not directly accessible, we contact the corresponding authors to obtain permission for research use. The detailed per-dataset descriptions are provided in the Appendix.

3.2 Data Standardization

The selected datasets differ in file organization, annotation formats, and metadata completeness. To ensure structural consistency, we standardize all data under a unified curation protocol. Original images stored in heterogeneous formats such as TIFF, JPEG, and PNG, with inconsistent naming conventions, are systematically reorganized. To prevent filename conflicts across sources, each image is assigned a unique identifier while preserving its source information. All images are converted to JPEG format at their original resolution. For each image, we construct a structured JSON record to store curated metadata. Key attributes such as biological domain, imaging modality, staining method, and host species are verified through original publications, communication with dataset providers, or consultation with domain experts. To further ensure reliability, all attributes are independently reviewed by two biologists, and samples with inconsistent judgments are removed.

3.3 Benchmark Tasks

Building on the established metadata, we introduce a suite of benchmark tasks to evaluate VLMs across different aspects of microscopic plant image understanding. The tasks are introduced as follows.

1) Modality Identification. This task focuses on identifying the microscopy modality used to acquire each image. PlantMicro includes three primary microscopy modalities: light microscopy, fluorescence microscopy, and electron microscopy. This task evaluates VLMs' ability to recognize modality-specific visual cues, such as bright-field illumination patterns in light microscopy, fluorescent signals in fluorescence microscopy, and fine surface textures in electron microscopy.

2) Domain Classification. This task aims to identify the biological domain of each image based on its visual characteristics. The images in PlantMicro are grouped into four major domains: Mycology, Nematology, Botany, and Cell Biology. The four domains correspond to fungal, nematode, plant tissue, and cellular imaging studies. This task evaluates a VLM's ability to understand different biological contexts in order to recognize domain-specific visual patterns such as spore morphology, nematode anatomy, or cellular structures.

3) Resin Identification. This task aims to identify the embedding resin used in electron microscopy sample preparation. Resin is used to embed and support biological samples so that they can be imaged and preserved without distortion, especially for electron microscopy and high-resolution light microscopy. Different resins provide varying levels of mechanical support and structural stability, which may influence image contrast, sharpness, and background consistency. PlantMicro includes samples embedded with commonly used resins such as Epon, Spurr, and Quetol-651. These materials lead to differences in imaging appearance, including contrast clarity, and and background uniformity. This task evaluates whether VLMs can capture these visual cues associated with different resins.

4) Stain/Dye Recognition. Stains and dyes are used to add contrast and specificity so microscopic structures become distinguishable. This task aims to identify the staining or fluorescent dye applied in the microscopic imaging process. PlantMicro contains samples labeled with various dyes such as DAPI, Green Fluorescent Protein (GFP), and Red Fluorescent Protein (RFP), which include both chemical and fluorescence staining. Different dyes result in distinct colors and signal distributions. This task evaluates the VLMs' ability to understand staining techniques and biological semantics purely from visual appearance.

5) Host Identification. In this task, VLMs are required to recognize the host organism shown in each microscopic image. The host refers to the plant species from which the sample is taken. Unlike macroscopic photos, microscopic images only show local cell or tissue structures without clear morphological contexts, making recognition difficult. Models need to use fine visual cues such as cell wall patterns, organelle arrangement, and color variations to infer which plant species the sample belongs to.

6) Pathogen Classification. This task involves fine-grained classification within each biological domain. In the fungal datasets, the goal is to distinguish different pathogenic species or infection types, while in the nematode datasets, it focuses on recognizing individual nematode species. These categories often exhibit high intra-class similarity and subtle inter-class differences in texture, shape, and structural patterns. The task evaluates whether models can capture fine-grained visual features to achieve accurate recognition within each domain.

7) Organelle Detection. The task assesses whether VLMs can robustly detect subcellular structures under heterogeneous visual conditions. Common targets include nuclei, mitochondria, plastids, and vacuoles. These structures differ in shape, texture, and contrast, but their appearances can vary across cell types, stains, and imaging modalities. The evaluation measures VLMs' capability to recognize organelles in microscopic images.

8) Organ Prediction. This task predicts which plant organ a microscopic image originates from, for example, leaf, root, cotyledon, or seed. It examines whether VLMs can relate cellular organization and tissue morphology to organ-level identity, capturing the structural hierarchy from microscopic features to macroscopic anatomy.

9) Target Counting. This task evaluates whether VLMs can accurately enumerate microscopic entities, such as spores or pollen grains, in densely populated visual scenes. Counting requires distinguishing true targets from visually similar debris and background artifacts, as well as separating tightly clustered instances, testing VLMs' ability for fine-grained visual parsing and quantitative reasoning. The average of objects per image is 8.2, and the median is 6.

10) Object Localization. This task evaluates the spatial grounding ability of VLMs by localizing a specified target and the prediction of its bounding box coordinates (*i.e.*, absolute pixel coordinates (x_1, y_1, x_2, y_2)).

Tasks 1-8 are evaluated under a multiple-choice setting, where models select the correct answer from four candidate options. For Tasks 9 and 10, we additionally evaluate models under a direct prediction setting, where models directly

generate the target count or bounding box coordinates in a predefined format. Predictions that violate the required format are treated as failures.

3.4 VQA Generation

Based on the defined benchmark tasks for PlantMicro, we construct a large collection of closed visual question answering (VQA) pairs to evaluate VLM performance across multiple aspects of microscopic image understanding. Each VQA pair follows a multiple-choice format, where models are required to select the correct answer from a set of candidate options. The questions and the ground truth answers are derived directly from the metadata associated with the microscopy images. For recognition tasks, distractor options are sampled from other valid values within the same metadata field to ensure semantic consistency and biological plausibility while avoiding visually implausible choices. For the counting task, distractor options are generated by perturbing the ground-truth count with nearby integer values to produce numerically plausible alternatives. For the localization task in the multiple-choice setting, candidate bounding boxes include the ground-truth box and several spatially perturbed alternatives with similar sizes but shifted positions. The candidate boxes are designed to be non-overlapping to ensure clear spatial separation among options. Since source microscopy datasets provide heterogeneous metadata fields (*e.g.*, some include "organelle" annotations while others do not, and some images contain multiple objects whereas others contain only one), the available question types differ across sources. Accordingly, the VQA pairs are generated by aligning each dataset with its valid metadata fields, ensuring that every task is derived only from images containing the required information. All generated VQA pairs are independently reviewed by two domain experts with backgrounds in plant pathology and microscopy to ensure reliability. The experts examine the correctness of the attribute–question alignment, the consistency between visual evidence and the ground-truth answer, and the biological plausibility of all candidate options. Samples with inconsistent judgments are excluded.

4 Experiments

4.1 Experiment Settings

Benchmarking VLMs. We evaluate a diverse set of VLMs that differ in architecture and size on the PlantMicro benchmark. These VLMs fall into two categories: closed-source models, including the GPT series [40], Gemini series [7], and open-source models, including CLIP [45], LLaVA [27], LLaVA-Next [22], and QwenVLM [4]. To ensure a fair comparison, all models are assessed under a unified evaluation pipeline that standardizes input processing and prompt design. We conduct experiments on a workstation with two NVIDIA A6000 GPUs.

Evaluation Metrics. We evaluate VLMs across three task categories with task-specific metrics. For multiple-choice identification tasks, we report both macro and micro accuracies. The macro and micro accuracy are defined as

Micro $= \sum_{i=1}^{C} TP_i / \sum_{i=1}^{C} N_i$, and Macro $= \frac{1}{C} \sum_{i=1}^{C} (TP_i/N_i)$, where C is the number of classes, TP_i is the number of correctly predicted samples for class i, and N_i is the total number of samples for class i. For counting tasks, models predict the number of target instances per image, and we report Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to quantify numerical deviation from ground truth. For localization tasks, models output bounding box coordinates, and performance is evaluated using mean Average Precision at an IoU (Intersection over Union) threshold of 0.5 (mAP@0.5) together with average IoU. To ensure consistent evaluation, all models are prompted with task-specific formats, including multiple-choice selection for identification, integer-only outputs for counting, and fixed-format bounding box predictions for localization. The predictions are regarded as incorrect if they violate the required format or differ from the ground-truth annotations.

4.2 Main Results

Multiple-choice Recognition Performance Analysis. Table 2 summarizes the main results, with the last row reporting the overall performance of all VLMs across recognition tasks. Performance variation across tasks reflects inherent differences in task difficulty. While modality, domain, resin, and stain identification largely rely on global visual cues, the remaining tasks demand fine-grained recognition of structurally diverse microscopic entities. VLMs achieve their strongest performance on the modality identification task. State-of-the-art closed-source models, such as GPT-5 and Gemini 2.5, surpass 95% accuracy, whereas open-source counterparts trail by a substantial margin, highlighting a pronounced capability gap in capturing global visual patterns. Both closed and open-source VLMs perform poorly on host and pathogen identification, with average accuracies around 30%. These results are marginally above random guessing, which suggests that VLMs still face considerable challenges in identifying host or pathogen types from microscopic evidence. For the remaining tasks, the average accuracies generally fall within the 40–50% range, among which stain recognition performs best. This may be attributed to the ability of VLMs to associate distinct color cues with specific stains, such as green fluorescence from GFP (green fluorescent protein). Overall, the results across tasks reveal that while VLMs are capable of coarse perceptual discrimination, they still fall short in fine-grained and biologically grounded understanding.

VLM Performance Comparison. Across most benchmark tasks, closed-source models consistently outperform open-source counterparts. Among them, Gemini 2.5-Flash achieves the highest overall performance, reaching a macro accuracy of 63.08%, highlighting the advantage of large proprietary models in visual–textual alignment under microscopic settings. Interestingly, GPT-5-mini slightly surpasses GPT-5 on several tasks, suggesting that smaller models may exhibit stronger robustness in certain fine-grained microscopic contexts. Open-source VLMs show greater performance variability. QwenVLM-7B attains the highest macro accuracy within this group (51.98%), approaching the lower range of

Table 2: Benchmark performance of vision–language models on PlantMicro. Results are provided for closed-source and open-source VLMs evaluated on identification tasks.

Model	Image-Level				Biological-Level				Overall	
	Mod.	Dom.	Res.	Sta.	Host	Path.	OrgL.	Org.	Macro	Micro
Random Choice	24.97	25.02	25.21	25.52	24.95	24.89	24.99	23.90	24.93	25.23
Closed-Source Vision Language Models(VLMs)										
GPT-4o mini [39]	73.23	62.97	51.85	50.03	30.78	29.00	49.26	62.96	51.26	53.34
GPT-5 mini [40]	**99.73**	57.54	50.62	52.51	31.40	**43.91**	50.74	71.60	57.26	61.17
GPT-5 [41]	97.72	55.46	51.85	58.46	34.93	42.71	49.07	64.19	56.79	60.56
Gemini 2.5-Flash-lite [7]	97.51	57.44	60.49	56.80	33.60	33.05	**56.48**	65.43	57.60	60.36
Gemini 2.5-Flash [7]	97.86	62.54	**85.19**	**61.45**	45.37	32.67	51.67	**67.90**	**63.08**	64.12
Gemini 2.5-Pro [7]	99.56	**63.86**	80.25	59.02	**46.82**	33.19	53.89	66.91	62.93	**65.07**
Avg (per task)	94.27	59.97	63.38	56.38	37.15	35.76	51.85	66.50	58.15	60.77
Open-Source Vision Language Models (VLMs)										
CLIP-ViT/32 [45]	67.37	33.05	9.64	48.51	31.70	**38.24**	26.85	16.05	33.93	42.90
CLIP-ViT/16 [45]	83.10	29.68	13.58	42.54	30.75	26.82	29.67	19.75	34.49	44.07
LLaVA-7B [27]	30.39	32.47	30.86	38.91	23.81	25.12	35.37	20.99	29.74	30.18
LLaVA-13B [27]	55.50	44.68	35.80	39.30	32.43	28.32	37.22	24.69	37.24	41.88
LLaVA-Next-7B [22]	29.11	34.49	32.10	50.45	23.08	19.55	30.19	49.38	33.54	30.57
LLaVA-Next-13B [22]	49.15	32.56	49.38	**49.97**	**40.10**	29.56	32.04	46.91	41.20	39.63
QwenVLM-7B [4]	**68.05**	**52.11**	**82.72**	49.68	30.75	25.21	**43.15**	**64.20**	**51.98**	**48.58**
Avg (per task)	54.67	37.01	36.30	45.62	30.37	27.55	33.50	34.57	37.45	39.69

closed-source performance. Models based on the LLaVA architecture show moderate performance, struggling with pathogen and host recognition. CLIP models perform substantially worse, particularly on resin and organ prediction tasks, where accuracies fall below the random baseline. This performance gap may be related to their contrastive pretraining on natural images, which limits adaptation to the visual patterns of microscopic domains. Overall, a consistent gap exceeding 10 percentage points remains between closed- and open-source models in both macro and micro accuracy, indicating that current open-source VLMs still face challenges when generalizing to specialized microscopic plant imagery without domain-specific adaptation.

Counting and Localization Performance. In addition to recognition tasks, we further evaluate VLMs on counting and localization under two evaluation settings: multiple-choice selection and direct prediction. In the multiple-choice setting, models select the correct answer from candidate options. In the direct prediction setting, models directly generate the target count or bounding box coordinates, which better reflects their ability to perform numerical reasoning and spatial grounding without predefined options. The results are presented in Table 3. From the multiple-choice results, closed-source models generally achieve higher counting accuracy than open-source models, suggesting stronger reasoning ability when candidate options are provided. GPT-5 achieves the highest multiple-choice counting accuracy, while QwenVLM-7B performs best among open-source models. In contrast, direct prediction reveals larger performance gaps. For counting, GPT-5 achieves the lowest MAE and RMSE, indicating more precise numerical estimation, whereas most open-source models exhibit substantially larger deviations from the ground truth. For localization, closed-source models do not consistently outperform open-source counterparts. The Gemini

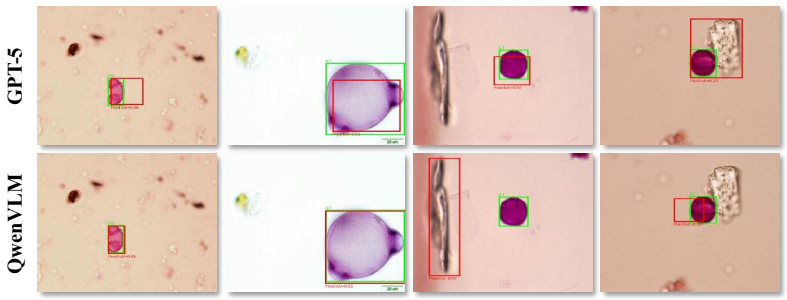


Fig. 4: Qualitative localization results of GPT-5 and QwenVLM-7B. Red boxes denote model predictions, while green boxes indicate ground-truth bounding boxes.

Table 3: Performance comparison on counting and localization tasks. Evaluation is conducted under two settings: multiple-choice selection and direct output prediction.

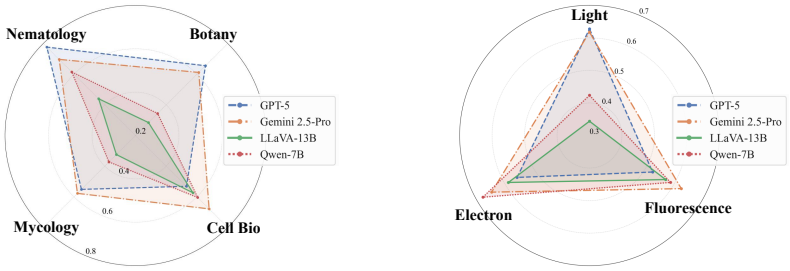
Model	Counting			Localization		
	MC Acc↑	MAE↓	RMSE↓	MC Acc↑	mAP@0.5↑	IoU↑
GPT-5-mini [40]	66.29	2.84	3.70	79.33	23.64	30.91
GPT-5 [41]	**71.34**	**2.07**	**3.02**	83.21	56.66	50.53
Gemini 2.5-Flash [7]	40.62	3.39	7.98	62.77	27.01	34.16
Gemini 2.5-Pro [7]	41.37	3.30	6.32	65.91	27.60	37.75
LLaVA-7B [27]	27.94	11.95	27.81	28.44	23.54	31.70
LLaVA-13B [27]	28.10	8.71	22.18	32.41	30.78	40.33
LLaVA-Next-7B [22]	28.76	8.67	22.75	31.98	28.10	34.48
LLaVA-Next-13B [22]	35.37	5.85	13.76	35.02	35.50	42.17
QwenVLM-7B [4]	46.45	3.38	9.24	**89.44**	**65.05**	**68.64**

series exhibits weak performance, whereas QwenVLM-7B surpasses GPT-5 and achieves the best spatial grounding results.

To further analyze this difference in spatial grounding ability, we visualize representative localization results from GPT-5 and QwenVLM-7B in Fig. 4. QwenVLM-7B produces bounding boxes that more tightly align with the target objects, indicating stronger spatial grounding. In contrast, GPT-5 roughly identifies object locations but produces less precise bounding boxes, while appearing less sensitive to background noise.

4.3 Performance across Domains and Modalities

To assess VLM robustness, we evaluate four representative models: GPT-5, Gemini 2.5-Pro, LLaVA-13B and QwenVLM-7B, across different biological domains. As demonstrated in Figure 5a, all models perform substantially better on nematology and cell biology, while botany and mycology remain more challenging. Specifically, GPT-5 attains the highest accuracy on nematology and botany, followed by Gemini 2.5-Pro, whereas LLaVA-13B and QwenVLM-7B lag notably behind. Gemini 2.5-Pro consistently surpasses QwenVLM-7B and LLaVA-13B across all domains but remains slightly weaker than GPT-5 on nematology. In contrast, GPT-5 exhibits the lowest accuracy on cell biology among all models. Figure 5b further summarizes performance across imaging modalities. Most



(a) Domain-wise comparison **(b)** Modality-wise comparison

Fig. 5: Accuracy comparison across biological domains and imaging modalities.

models achieve their highest accuracy on electron micrographs, as the strong performance on electron images may be attributed to their more uniform and structured appearance. Fluorescence images provide the second-best results, and light microscopy generally yields the lowest overall performance. However, GPT-5 shows a different pattern: it achieves the best performance on light microscopy, which is the most challenging modality for open-source VLMs. This suggests that GPT-5 is better adapted to light microscopy images.

4.4 Analysis of Reasoning and In-Context Prompting

Effect of CoT prompting. Chain-of-Thought (CoT) prompting [57] has shown strong effectiveness in arithmetic, commonsense, and symbolic reasoning tasks. To examine whether explicit reasoning guidance can similarly benefit microscopic understanding, we evaluate the impact of CoT prompting on identification tasks. Specifically, we adopt a three-step CoT setting following [20], where a structured prompting scheme instructs the model to follow an observation–reasoning–selection workflow when answering each question. We conduct experiments using Gemini 2.5-Pro and QwenVLM-7B, the best-performing closed-source and open-source models in multiple-choice identification tasks, to ensure a stable and representative evaluation. Results with and without CoT are summarized in Table 4. Under the CoT setting, Gemini 2.5-Pro shows slight improvements on the Stain and Organ tasks but exhibits performance drops on most other tasks, leading to a lower overall accuracy. In contrast, QwenVLM-7B achieves noticeable gains across multiple tasks, suggesting that structured reasoning guidance may better support open-source models, which tend to rely more on explicit reasoning steps when handling microscopic cues.

Impact of In-Context Prompting. We evaluate the effect of in-context prompting in microscopic scenarios using a 1-shot setting with the best-performing models, GPT-5 and QwenVLM-7B, on both counting and localization tasks. Results in Table 5 show that introducing a single in-context example does not lead to consistent performance improvements. Specifically, QwenVLM-7B shows a slight improvement in counting but degraded performance on localization, while GPT-5 exhibits performance drops on both tasks. We attribute this degradation to

Table 4: Effect of CoT prompting on Gemini 2.5-Pro and QwenVLM-7B.

Model	CoT	Image-Level				Biological-Level				Overall	
		Mod.	Dom.	Res.	Sta.	Host	Path.	OrgL.	Org.	Macro	Micro
Gemini 2.5-Pro [7]	✗	**99.56**	**63.86**	**80.25**	59.02	**46.82**	**33.19**	53.89	66.91	62.93	65.07
	✓	98.80	55.69	66.66	**62.14**	35.31	32.89	**59.81**	**79.01**	61.28	61.44
QwenVLM-7B [4]	✗	68.05	52.11	82.72	49.68	30.75	25.21	43.15	64.20	51.98	48.58
	✓	**91.92**	**57.28**	**85.19**	**50.52**	**37.47**	**25.26**	**49.63**	**66.67**	**57.99**	**57.72**

Table 5: Few-shot in-context calibration results on counting and localization tasks.

Model	Setting	Counting			Localization		
		MC Acc↑	MAE↓	RMSE↓	MC Acc↑	mAP@0.5↑	IoU↑
GPT-5 [41]	0-shot	**71.34**	**2.07**	**3.02**	**83.21**	**56.66**	**50.53**
GPT-5 [41]	1-shot	68.29	2.28	4.13	68.73	53.73	43.34
QwenVLM-7B [4]	0-shot	46.45	3.38	9.24	**89.44**	**65.05**	**68.64**
QwenVLM-7B [4]	1-shot	**47.18**	**3.11**	**6.73**	72.76	62.79	65.85

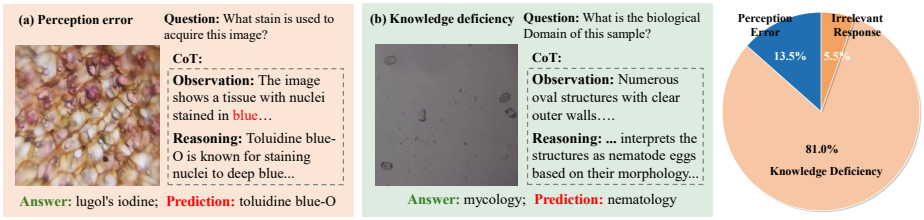
misalignment between the in-context demonstration and the target image. Due to the high visual similarity, repetitive structures, and dense object distributions in microscopic images, models may over-rely on patterns from the demonstration rather than grounding predictions on the input image itself.

4.5 Error Analysis

As VLMs exhibit limited capability on PlantMicro, we conduct an error analysis to identify their major failure modes. For each benchmark task, we randomly select 25 erroneous samples and trace the CoT reasoning process of Gemini 2.5-Pro to determine at which stage the model deviates from the correct reasoning path. The observed errors can be grouped into three categories: perception errors, knowledge deficiencies, and irrelevant responses. In perception errors, the model fails at the earliest stage of CoT reasoning: visual observation. Figure 6a presents an example where the stain color is misidentified as blue, which subsequently leads the model to predict toluidine blue O instead of the ground-truth label. In contrast, knowledge deficiencies arise when the model correctly perceives the visual evidence but lacks the biological knowledge needed for accurate interpretation. As demonstrated in Figure 6b, although the model correctly identifies the oval structure of spores, it mistakenly links them to nematode eggs, indicating that the model has difficulty distinguishing entities at a fine-grained level. These two error types correspond to the observation and reasoning stages in the CoT process, respectively. Irrelevant responses, by contrast, refer to cases in which the model produces semantically unrelated outputs. According to the error distribution shown in Figure 6c, knowledge deficiency accounts for the majority of failures, indicating that current VLMs have not yet been effectively adapted to the plant microscopy domain.

4.6 Investigation on RAG

To examine whether domain-specific knowledge can improve VLM performance in microscopic scenarios, we introduce a retrieval-augmented generation (RAG)



(a) Perception error. **(b)** Knowledge deficiency. **(c)** Error types.

Fig. 6: Error analysis. (a) Example of perception errors. (b) Example of knowledge deficiencies. (c) Distribution of error types across tasks.

Table 6: Impact of RAG on microscopic VQA tasks. † denotes methods using RAG.

Model	Image-Level				Biological-Level				Overall	
	Mod.	Dom.	Res.	Sta.	Host	Path.	OrgL.	Org.	Macro	Micro
LLaVA-13B [27]	55.50	44.68	35.80	39.30	32.43	28.32	37.22	24.69	37.24	41.88
LLaVA-13B† [27]	**86.17**	**52.73**	**85.19**	**52.66**	**54.57**	**55.43**	**43.15**	**32.10**	**57.75**	**60.21**
LLaVA-Next-13B [22]	49.15	32.56	49.38	**49.97**	40.10	29.56	32.04	46.91	41.20	39.63
LLaVA-Next-13B† [22]	**90.28**	**48.90**	**80.25**	44.49	**50.91**	**57.51**	**36.30**	**56.79**	**58.18**	**60.05**
QwenVLM-7B [4]	68.05	52.11	82.72	49.68	30.75	25.21	43.15	64.20	51.98	48.58
QwenVLM-7B† [4]	**98.15**	**58.59**	**83.95**	**81.42**	**56.75**	**39.69**	**62.08**	**68.41**	**68.63**	**69.04**

setting for open-source VLMs (marked with †). Specifically, we retrieve the most visually similar example from the benchmark pool based on visual embedding similarity and prepend it as an in-context demonstration. To ensure task consistency, retrieval is restricted to samples within the same task category. As shown in Table 6, incorporating RAG consistently improves performance across most microscopic VQA tasks. These results suggest that the performance bottlenecks of current VLMs largely stem from insufficient domain-specific knowledge.

5 Conclusion

In this paper, we introduce PlantMicro, a comprehensive benchmark for evaluating vision–language models in microscopic plant image understanding. By integrating diverse microscopy datasets spanning multiple biological domains and imaging modalities, PlantMicro enables systematic evaluation across recognition, counting, and localization tasks. Extensive experiments demonstrate that while state-of-the-art VLMs such as Gemini 2.5 and GPT-5 can achieve strong performance in image-level modality identification, they exhibit substantial degradation in fine-grained biological entity recognition, spatial localization, and object counting. Our error analysis indicates that these limitations primarily arise from insufficient domain-specific knowledge in plant microscopy. Further experiments show that incorporating retrieval-augmented generation (RAG) consistently improves performance, highlighting the potential of knowledge-enhanced approaches for microscopic visual understanding. Overall, PlantMicro bridges the gap between general-purpose VLMs and domain-specific microscopic understanding, providing a foundation for future research in plant microscopy.

Acknowledgement

This work was partially supported by the Grains Research and Development Corporation (GRDC) through the Analytics for the Australian Grains Industry (AAGI) Strategic Partnership (UOQ2301-010OPX) and the project Exploring Data Efficient Machine Learning for Australian Grains Disease Recognition (USQ2605-002BGX), and by The University of Queensland (DVCR2201A).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Amrani, A., Sohel, F., Diepeveen, D., Murray, D., Jones, M.G.: Deep learning-based detection of aphid colonies on plants from a reconstructed brassica image dataset. Computers and electronics in agriculture **205**, 107587 (2023)
3. 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. IEEE (2025)
4. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
5. Biswas, S., Barma, S.: A large-scale optical microscopy image dataset of potato tuber for deep learning based plant cell assessment. Scientific Data **7**(1), 371 (2020)
6. Chakrabarty, S., Shashank, P.R., Deb, C.K., Haque, M.A., Thakur, P., Kamil, D., Marwaha, S., Dhillon, M.K.: Deep learning-based accurate detection of insects and damage in cruciferous crops using yolov5. Smart Agricultural Technology **9**, 100663 (2024)
7. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
8. Crespo-Michel, A., Alonso-Arévalo, M.A., Hernández-Martínez, R.: Developing a microscope image dataset for fungal spore classification in grapevine using deep learning. Journal of Agriculture and Food Research **14**, 100805 (2023)
9. Dongre, V., Gui, C., Garg, S., Nayyeri, H., Tur, G., Hakkani-Tür, D., Adve, V.S.: Mirage: A benchmark for multimodal information-seeking and reasoning in agricultural expert-guided conversations. arXiv preprint arXiv:2506.20100 (2025)
10. Du, B., Haensch, R., Alfarraj, S., Rennenberg, H.: Strategies of plants to overcome abiotic and biotic stresses. Biological Reviews **99**(4), 1524–1536 (2024)
11. Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P.: Primary production of the biosphere: integrating terrestrial and oceanic components. science **281**(5374), 237–240 (1998)
12. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)

13. Gauba, A., Pi, I., Man, Y., Pang, Z., Adve, V.S., Wang, Y.X.: Agmmu: A comprehensive agricultural multimodal understanding and reasoning benchmark. In: arXiv preprint arXiv:2504.10568 (2025)
14. Gong, H., Xiao, L., Wang, X.: Segmentation and coverage measurement of maize canopy images for variable-rate fertilization using the mcac-unet model. Agronomy **14**(7), 1565 (2024)
15. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
16. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
17. Indarti, S., Shabrina, N.H., Maharani, R.: Microscopic image dataset of plant-parasitic nematode. Data in Brief p. 111687 (2025)
18. Iqbal, Z., Iqbal, M.S., Hashem, A., Abd_Allah, E.F., Ansari, M.I.: Plant defense responses to biotic stress and its interplay with fluctuating dark/light conditions. Frontiers in Plant Science **12**, 631810 (2021)
19. Javidan, S.M., Banakar, A., Vakilian, K.A., Ampatzidis, Y., Rahnama, K.: Diagnosing the spores of tomato fungal diseases using microscopic image processing and machine learning. Multimedia Tools and Applications **83**(26), 67283–67301 (2024)
20. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
21. Lämke, J., Bäurle, I.: Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. Genome biology **18**(1), 124 (2017)
22. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
23. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
24. Li, Y., Che, Y., Zhang, H., Zhang, S., Zheng, L., Ma, X., Xi, L., Xiong, S.: Wheat growth stage identification method based on multimodal data. European Journal of Agronomy **162**, 127423 (2025)
25. Lim, J.S., Luo, Y., Chen, Z., Wei, T., Chapman, S., Huang, Z.: Track any peppers: Weakly supervised sweet pepper tracking using vlms. arXiv preprint arXiv:2411.06702 (2024)
26. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536. Springer (2023)
27. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
28. Liu, K.H., Yang, M.H., Huang, S.T., Lin, C.: Plant species classification based on hyperspectral imaging via a lightweight convolutional neural network model. Frontiers in Plant Science **13**, 855660 (2022)
29. Liu, K.H., Yang, M.H., Huang, S.T., Lin, C.: Plant species classification based on hyperspectral imaging via a lightweight convolutional neural network model. Frontiers in Plant Science **13**, 855660 (2022)

30. Liu, X., Liu, Z., Hu, H., Chen, Z., Wang, K., Wang, K., Lian, S.: A multimodal benchmark dataset and model for crop disease diagnosis. In: European Conference on Computer Vision. pp. 157–170. Springer (2024)
31. Lozano, A., Nirschl, J.J., Burgess, J., Gupte, S.R., Zhang, Y., Unell, A., Yeung-Levy, S.: Micro-bench: A microscopy benchmark for vision-language understanding. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track
32. Lu, Y., Du, S., Ji, Z., Yin, X., Jia, W.: Odl net: Object detection and location network for small pears around the thinning period. Computers and Electronics in Agriculture **212**, 108115 (2023)
33. Lück, S., Bourras, S., Douchkov, D.: Deep phenotyping platform for microscopic plant-pathogen interactions. Frontiers in Plant Science **16**, 1462694 (2025)
34. Mano, S., Miwa, T., Nishikawa, S.i., Mimura, T., Nishimura, M.: The plant organelles database (podb): a collection of visualized plant organelles and protocols for plant organelle research. Nucleic acids research **36**(suppl_1), D929–D937 (2007)
35. Mano, S., Nakamura, T., Kondo, M., Miwa, T., Nishikawa, S.i., Mimura, T., Nagatani, A., Nishimura, M.: The plant organelles database 3 (podb3) update 2014: integrating electron micrographs and new options for plant organelle research. Plant and Cell Physiology **55**(1), e1–e1 (2014)
36. Meyerowitz, E.M.: Genetic control of cell division patterns in developing plants. Cell **88**(3), 299–308 (1997)
37. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Frontiers in plant science **7**, 215232 (2016)
38. Ni, X., Wang, F., Huang, H., Wang, L., Wen, C., Chen, D.: A cnn-and self-attention-based maize growth stage recognition method and platform from uav orthophoto images. Remote Sensing **16**(14), 2672 (2024)
39. OpenAI: Gpt-4o mini (2024), `https://platform.openai.com/docs/models#gpt-4o-mini`, accessed: 2025-11-11
40. OpenAI: Gpt-5 mini (2025), `https://platform.openai.com/docs/models/gpt-5-mini`, accessed: 2025-11-11
41. OpenAI: Introducing gpt-5 (August 2025)
42. Ovečka, M., von Wangenheim, D., Tomančák, P., Šamajová, O., Komis, G., Šamaj, J.: Multiscale imaging of plant development by light-sheet fluorescence microscopy. Nature plants **4**(9), 639–650 (2018)
43. Pawlak, K., Kołodziejczak, M.: The role of agriculture in ensuring food security in developing countries: Considerations in the context of the problem of sustainable food production. Sustainability **12**(13), 5488 (2020)
44. Prunet, N., Duncan, K.: Imaging flowers: a guide to current microscopy and tomography techniques to study flower development. Journal of experimental botany **71**(10), 2898–2909 (2020)
45. 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. PmLR (2021)
46. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
47. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)

48. Rufián, J.S., Macho, A.P., Corry, D.S., Mansfield, J.W., Ruiz-Albert, J., Arnold, D.L., Beuzón, C.R.: Confocal microscopy reveals in planta dynamic interactions between pathogenic, avirulent and non-pathogenic pseudomonas syringae strains. Molecular plant pathology **19**(3), 537–551 (2018)
49. Shinoda, R., Inoue, N., Kataoka, H., Onishi, M., Ushiku, Y.: Agrobench: Vision-language model benchmark in agriculture. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7634–7644 (2025)
50. 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, pp. 249–253 (2020)
51. Takehara, S., Takaku, Y., Shimomura, M., Hariyama, T.: Imaging dataset of fresh hydrous plants obtained by field-emission scanning electron microscopy conducted using a protective nanosuit. PloS one **15**(5), e0232992 (2020)
52. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
53. Wang, A., Zhang, W., Wei, X.: A review on weed detection using ground-based machine vision and image processing techniques. Computers and electronics in agriculture **158**, 226–240 (2019)
54. Wang, D., Wang, J., Li, W., Guan, P.: T-cnn: Trilinear convolutional neural networks model for visual detection of plant diseases. Computers and Electronics in Agriculture **190**, 106468 (2021)
55. Wang, H., Guan, Y., Meng, F., Zhao, C., Yan, L., Yang, Y., Jiang, J.: Agri-cm3: A chinese massive multi-modal, multi-level benchmark for agricultural understanding and reasoning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11729–11754 (2025)
56. Wang, Z., Zenkl, R., Greche, L., De Solan, B., Bernigaud Samatan, L., Ouahid, S., Visioni, A., Robles-Zazueta, C.A., Pinto, F., Perez-Olivera, I., et al.: The global wheat full semantic organ segmentation (gwfss) dataset. bioRxiv pp. 2025–03 (2025)
57. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
58. Wei, T., Chen, Z., Huang, Z., Yu, X.: Benchmarking in-the-wild multimodal disease recognition and a versatile baseline. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 1593–1601 (2024)
59. Wei, T., Chen, Z., Yu, X.: Snap and diagnose: An advanced multimodal retrieval system for identifying plant diseases in the wild. In: Proceedings of the 6th ACM International Conference on Multimedia in Asia. pp. 1–3 (2024)
60. Wei, T., Chen, Z., Yu, X., Chapman, S., Melloy, P., Huang, Z.: Plantseg: A large-scale in-the-wild dataset for plant disease segmentation. arXiv preprint arXiv:2409.04038 (2024)
61. Wei, T., Yu, X., Chen, Z., Chapman, S., Huang, Z.: Augment to segment: Tackling pixel-level imbalance in wheat disease and pest segmentation. arXiv preprint arXiv:2509.09961 (2025)
62. Yordanov, M., d'Andrimont, R., Martinez-Sanchez, L., Lemoine, G., Fasbender, D., Van der Velde, M.: Crop identification using deep learning on lucas crop cover photos. Sensors **23**(14), 6298 (2023)
63. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. pp. 18123–18133 (2022)
64. Zhang, J., Min, A., Steffenson, B.J., Su, W.H., Hirsch, C.D., Anderson, J., Wei, J., Ma, Q., Yang, C.: Wheat-net: An automatic dense wheat spike segmentation method based on an optimized hybrid task cascade model. Frontiers in Plant Science **13**, 834938 (2022)
65. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
66. Zhong, L., Hu, L., Zhou, H.: Deep learning based multi-temporal crop classification. Remote sensing of environment **221**, 430–443 (2019)
67. Zhu, X., Chen, F., Qiao, C., Zhang, Y., Zhang, L., Gao, W., Wang, Y.: Cucumber pathogenic spores' detection using the gcs-yolov8 network with microscopic images in natural scenes. Plant Methods **20**(1), 131 (2024)
68. Zu, B., Cao, T., Li, Y., Li, J., Ju, F., Wang, H.: Swint-srnet: Swin transformer with image super-resolution reconstruction network for pollen images classification. Engineering Applications of Artificial Intelligence **133**, 108041 (2024)