AgroBench: Vision-Language Model Benchmark in Agriculture

Risa Shinoda[1,2,4], Nakamasa Inoue[3,4], Hirokatsu Kataoka[4,5], Masaki Onishi[4], Yoshitaka Ushiku[6]

[1]The University of Osaka [2]Kyoto University [3]Institute of Science Tokyo
[4]National Institute of Advanced Industrial Science and Technology (AIST)
[5]Visual Geometry Group, University of Oxford [6]OMRON SINIC X

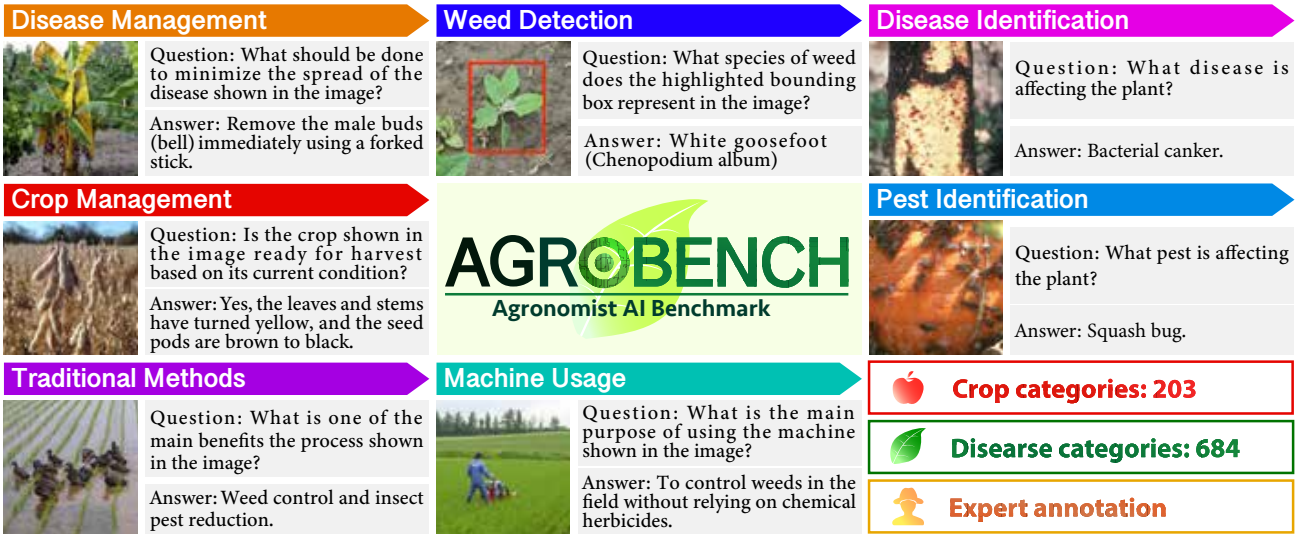

Figure 1. We present AgroBench (**Agro**nomist AI **Bench**mark) designed to comprehensively evaluate 682 disease categories across 203 agricultural crop types for 7 vision-language question-answer tasks. In the era of larger-scale vision-language models (VLMs), our AgroBench is obviously non-trivial in terms of many more crop and disease categories with all expert annotations for establishing QA benchmarks in the agricultural domain.

Abstract

*Precise automated understanding of agricultural tasks such as disease identification is essential for sustainable crop production. Recent advances in vision-language models (VLMs) are expected to further expand the range of agricultural tasks by facilitating human-model interaction through easy, text-based communication. Here, we introduce AgroBench (**Agro**nomist AI **Bench**mark), a benchmark for evaluating VLM models across seven agricultural topics, covering key areas in agricultural engineering and relevant to real-world farming. Unlike recent agricultural VLM benchmarks, AgroBench is annotated by expert agronomists. Our AgroBench covers a state-of-the-art range of categories, including 203 crop categories and 682 disease categories, to thoroughly evaluate VLM capabilities. In our evaluation on AgroBench, we reveal that VLMs have room for improvement in fine-grained identification tasks. Notably, in weed identification, most open-source VLMs perform close to random. With our wide range of topics and expert-annotated categories, we analyze the types of errors made by VLMs and suggest potential pathways for future VLM development. Our dataset and code are available at* https://dahlian00.github.io/AgroBenchPage/.

1. Introduction

Agriculture is a fundamental process for humans to produce crops to live and stay healthy. With the development of computer vision technology, effective and automated management of external crop factors such as diseases and pests has been explored, contributing to stable crop

Dataset	Annot.	Crop	Weed	Disease	Pest	Images	QA pairs	Main Purpose
Agri-LLaVA [48]	GPT-4	29	-	109	112	391k	391k (Synthetic)	Training
AgroInstruct [4]	GPT-4	174	4	74	12	70k	70k (Synthetic)	Training
CDDM [20]	GPT-4	15	–	60	–	137k	1M (Synthetic)	Training
AgroBench (Ours)	**Expert**	**203**	**108**	**682**	**134**	3,745	4,342 (**Expert**)	Evaluation

Table 1. **Comparison of agricultural vision datasets.** Expert refers to 'human expert' in this context. AgroBench provides a comprehensive evaluation framework for assessing VLMs, featuring multiple tasks and a wide range of categories.

production. This includes detecting and classifying undesirable conditions like diseases [1, 10, 13, 34, 38, 41, 47] and pests [3, 49, 53], as well as general plant management tasks such as crop classification [52, 54] and recognition of crop maturity [22, 25, 37, 39], and structure understanding [21, 43].

To maintain stable crop production, it is important to recognize a wide range of crop conditions, including undesirable situations such as diseases and pests, and to know how to respond appropriately. A single visual model that can handle various conditions, not only disease and pest detection but also treatment and general crop management, would be highly beneficial. However, most existing approaches use task-specific models. These models usually require large amounts of training images and manual annotations for each task. As a result, farmers often need to use several different models depending on the situation. This increases complexity and makes the overall system less accessible for practical use in agriculture.

For general purpose visual tasks, vision-language models (VLMs) [17, 19, 32, 33, 45, 46, 50] have become widespread recently because they can understand task definitions provided by natural language prompts, covering a wide range of applications without the need for task-specific model training. The recognition ability of VLMs is closely connected to image recognition itself and linked to an object's words, and supports open-vocabulary recognition through web-scale training. Here, zero-shot recognition and few-shot adaptation have also been realized by VLMs with language representations. This ability opens up a wide range of applications; therefore, we believe this can be applied to agricultural scenarios as well. It is worth noting that VLMs offer an easy-to-use interface for the general public, especially the question-answer (QA) and conversation modes. To investigate the range of tasks that VLMs obtained through large-scale vision-language training can cover, recent studies have introduced various benchmark datasets to fully evaluate the VLMs, including tasks such as understanding diagrams and charts [26, 28], and question answering requiring specialized knowledge [7, 55, 56].

However, VLM research remains underexplored in agriculture due to a lack of benchmark datasets that include diverse tasks and categories in agriculture. Several pioneering works [20, 48] have adapted open-source VLMs such as LLaVA [18] to the agricultural domain by fine-tuning them on synthetic datasets generated from closed-source VLMs such as GPT-4o. This is recognized as an effective approach because black-box VLMs often possess a certain amount of agricultural knowledge obtained from the Internet. While this helps in generating responses that are generally acceptable to experts, there is almost no way to verify whether the answers are truly correct. Moreover, in the limited evaluation of categories, we are insufficient for fully assessing VLM knowledge; whether they can answer a wide range of types, such as diseases, pests, and weeds. These limitations motivated us to develop a benchmark dataset in order to evaluate the VLM's broad knowledge in the agricultural domain and its applicability as a practical application.

Here, we introduce AgroBench (**Agro**nomist AI **Bench**mark), a comprehensive benchmark dataset for VLM for the agricultural domain, covering a state-of-the-art range of categories for agriculture-focused benchmark datasets for VLM; 682 disease, 134 pest, 108 weed, and 203 crop categories (Figure 1). We cover not only identification tasks but also crop production and disease management knowledge. Moreover, we include other important topics related to crop management with 98 machine categories and 77 traditional management methods to investigate more about VLM's ability. We carefully selected benchmark tasks from key agricultural engineering research areas, and also the tasks that address challenges faced by farmers in real agricultural scenarios. We employ a multiple-choice format, and all questions are annotated by human agronomist experts, which overcomes the limitations of previous synthetically created datasets. We carefully selected images from Creative Commons-licensed and publicly available datasets, including real farm conditions. Also, under our manual annotation process, unclear images were removed. As shown in Table 1, AgroBench contains data with the most diverse disease and crop variation categories and many more crop/weed annotations.

Our main contributions are as follows;

- We developed AgroBench, a benchmark dataset for VLM, to assess their broad agricultural knowledge of VLMs and evaluate their applicability in practical applications.

- In our AgroBench evaluation, VLMs tend to achieve high performance in disease and crop management tasks; however, there is still room for improvement in weed and disease identification.
- Our AgroBench enables error analysis by providing broader category annotations, highlighting future directions for model training focus areas.

2. Related Work

2.1. Computer Vision for Agriculture

With the rise of computer vision techniques driven by advancements in deep learning, a wide range of agricultural tasks have been explored over the past decade. In particular, dataset construction and benchmarking play an important role in developing models and practical applications in agriculture, while also paving the way for forward-looking advancements. Disease identification is a key research focus across various crops, such as rice [1, 34], tomato [10], cacao [13], and sugarcane [47]. Multicrop datasets have also been created [29, 41]. PlantDoc [41] covers 13 species and 17 classes focused on leaf-based diseases. Plant Village [29] offers 39 classes containing both diseased and healthy leaf categories. In addition to crop disease identification, pest identification [49, 53], weed identification [12, 31, 42] have also been studied.

While most of the agricultural datasets primarily focus on visual data, few computer vision studies in agriculture investigate the potential of multi-modal approaches. The PlantWild dataset [52] includes 56 plant-disease class pairs, which are collected through image search engines. They implement a CLIP-based model and show the possibility of training with combined text and image data. In an instruction-based format, the CDDM [20] has created 16 categories of crops and 60 categories of crop diseases dataset, which generates instructional data using GPT. While these synthetically created datasets explore the agriculture multi-modal models, there remains a lack of datasets for comprehensive multi-modal model evaluation, validated by human experts and covering a wide range of tasks and categories.

2.2. Vision-Language Models

Models. Visual models in the computer vision field have been accelerated by language modality with the data resource of sentence-level inputs and web-scale texts. Specifically, CLIP [35] has made a significant contribution in this context, by text and image feature alignment through contrastive learning. CLIP played a key role in introducing more sophisticated visual representation into language explanations e.g., Flamingo [2], BLIP [14, 15], Qwen [5], PaLI [8], LLaVA [17, 18], CogVLM [50], and Emu [44, 45, 51]. Closed-source VLMs achieve state-of-the-art performance, such as GPT-4o [32] and Gemini Pro [46], which are said to acquire human-level knowledge across diverse fields on the Internet.

Benchmarks. Along this line, several recent studies introduced benchmark datasets. Especially, the representative examples in terms of universal knowledge benchmarking include MMMU [55], MMMU-Pro [56] and MMStar [6] for multi-modal understanding. These benchmarks contain highly diverse domains (e.g., natural, graph, illustration, medical images) and academic fields (e.g., science, engineering, art, and medical fields) under the tasks of vision language such as question-answering and reasoning. The series of MMMUs have been verified with foundation models like GPT-4V, but they cannot answer the questions perfectly. More specific domain datasets have also been proposed, such as those for Medicine [27, 36], Chart [26, 28, 40], and Video understandings [9, 16, 30], contributing to the accurate evaluation of VLMs and guiding future research directions.

3. AgroBench

This section introduces AgroBench, the first comprehensive benchmark dataset to evaluate VLM models from the perspective of agricultural vision tasks. The benchmark consists of seven tasks covering a wide range of tasks selected from key agricultural engineering research areas, as well as tasks that address real-world challenges faced by farmers in real agriculture scenarios. AgroBench includes 682 disease categories, 134 pest categories, 203 crop categories, and 108 weed categories, representing the largest number of categories in each area to date, to the best of our knowledge.

3.1. Benchmark Tasks in Agricultural Scenes

The seven benchmark tasks encompass key research areas in agricultural engineering as well as real-world challenges faced by farmers. To facilitate VLM evaluation, we also provide prompts to address each task in a question-answer format. The details of each task are described below.

1) Disease Identification (DID). The DID task aims to accurately diagnose and classify crop diseases. This is a key task in agriculture to protect crop health and maximize yields. Our benchmark provides 1,502 QA pairs that cover 370 disease categories, 160 crop categories and 682 crop-disease combinations. To thoroughly evaluate the capabilities of VLMs, we include four misleading disease labels for each image, featuring diseases with similar symptoms or common diseases affecting the target plant. VLMs are required to diagnose the disease based on the image, considering both symptoms and the crop species. Figure 2a illustrates example images.

2) Pest Identification (PID). The PID task aims to identify pests to prevent infestations that can severely impact crop health. Accurate identification reduces economic losses and

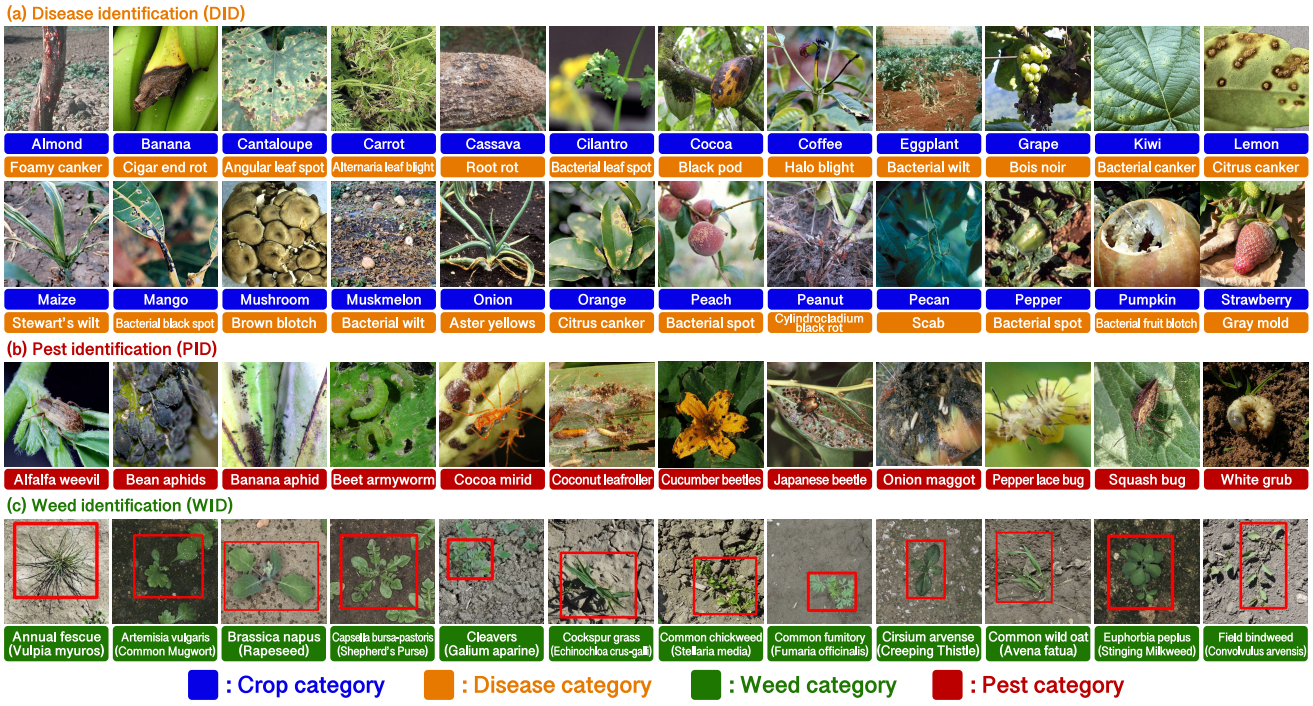

Figure 2. **Examples of labeled images for DID, PID, and WID tasks.** Our dataset includes 682 crop-disease pairs, 134 pest categories, and 108 weed categories. We prioritized collecting images from real farm settings.

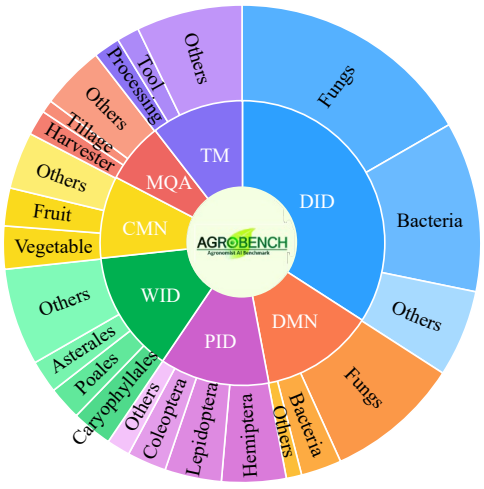

Figure 3. **Seven benchmark tasks in AgroBench.** AgroBench includes multiple topics with a diverse range of categories. The total accuracy is calculated by the average of each task to mitigate the difference in QAs.

minimizes harm to the environment. Our benchmark provides 544 labeled images that cover 134 pest categories, including insects, mites, and other organisms harmful to plants. For categories where we obtain multiple images, we select multiple insect growth stages whenever possible. To fully evaluate the VLMs, we assign alternative choices that closely resemble or are commonly associated with the target crops. Figure 2b shows examples of labeled pest images.

3) Weed Identification (WID). The WID task aims to identify weed species. Our benchmark includes 609 images of weeds with ground-truth bounding boxes, covering 108 weed species commonly found in farm fields. We assign bounding boxes because multiple weed types often grow closely together, and we want to clarify which one is the target. Specifically, VLMs are required to identify the weed species within the provided bounding box on the image. Figure 2c shows examples of labeled weed images.

4) Crop Management (CMN). Crop management focuses on optimizing farming practices to facilitate crop growth. This involves making decisions on irrigation, fertilization, planting times, and other cultivation practices. Our benchmark provides 411 question-answer pairs for this task. VLMs are required to analyze images of crops and recommend appropriate management strategies by considering factors such as crop health, growth stage, and environmental conditions visible in the images given five answer candidates. In Figure 4a and b, we show two examples of the harvest timing for white and green asparagus, respectively. Our dataset includes complex questions that take into account the differences in their harvest timing.

5) Disease Management (DMN). Disease management aims to control and reduce diseases in crops. This involves informed decisions on interventions such as applying pesticides, adopting resistant crop varieties, or modifying cul-

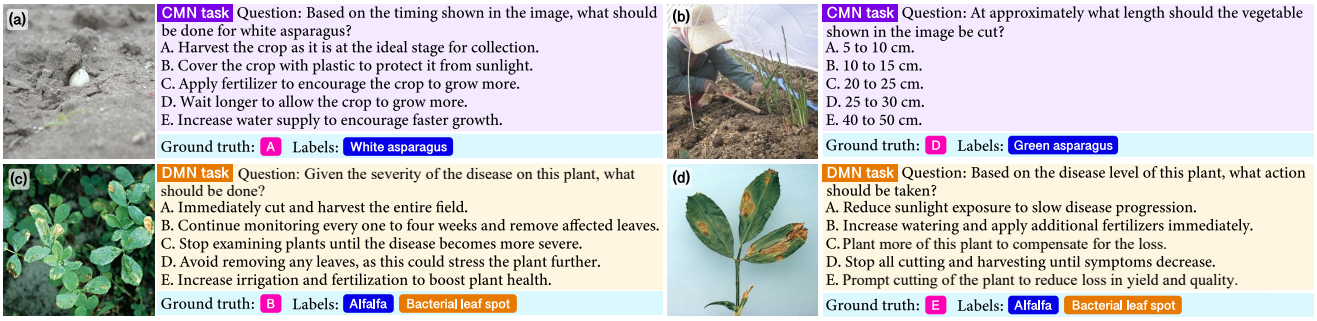

Figure 4. **Examples of QA pairs for CMN and DMN tasks.** (a) and (b) Crop management QA types for white asparagus and asparagus, respectively. Their difference in the harvest timing affects the answer's difference correctly. (c) and (d) Disease management QA types for the alfalfa bacterial leaf spot with the initial and severe symptoms, respectively. Based on the severity of the symptoms, the annotator changes the answer.

tivation practices. Our benchmark provides 569 question-answer pairs covering 141 crop-disease combination categories. Since many diseases share the same management strategies regardless of the crop, we carefully select a diverse range of disease types and management strategies. VLMs are required not only to identify the disease but also to recommend appropriate management strategies based on images of affected crops. In Figure 4c and d show example QAs from the Disease Management tasks. Both images depict Bacterial leaf spot on alfalfa, with c showing the initial stage of the disease and d showing the severe stage. Based on the severity of the disease, we provide different answer options and set different correct answers. The input prompt consists of a question and five answer candidates.

6) Machine Usage QA (MQA). Machine usage QA addresses the correct use and choice of agricultural machinery depending on the task and farming conditions. Selecting the appropriate machinery is essential for efficient farming. Our benchmark provides 303 question-answer pairs covering 98 machine categories. Given that many crops share the same machinery (e.g., soil preparation and irrigation), the coverage of this category is comprehensive. VLMs are required to answer questions about machinery operation or select the appropriate machine for a given scenario based on images of machinery or farming conditions.

7) Traditional Management (TM). Traditional management methods involve natural and sustainable approaches to farming, such as the use of organic fertilizers, terrace farming, and agroforestry. Recent computer vision studies have not focused on these traditional practices, although many are still used by certain local farmers. Our benchmark includes 404 question-answer pairs, including 77 traditional management practices. VLMs are required to identify the management method or explain its effectiveness, given five answer choices.

3.2. Dataset Construction

Image Selection. To establish a high-quality benchmark covering a wide range of crops, diseases, and pest categories we initially curated around 50,000 agricultural images from websites supervised by plant pathologists, either where redistribution was permitted or where we obtained redistribution permission. We selected images in real farm settings as much as possible, as shown in Figure 2, to evaluate real-world scenarios. When licensed images for a target category were limited, we used laboratory setting images. For curation, we obtained images along with their corresponding labels. The annotator is one of the authors, who holds a Ph.D. in Agriculture. In the human evaluation conducted during the experiment, other individuals with a Ph.D. or M.S. degree in Agriculture reviewed questions to assess the quality of the QA pairs. The annotator selected images that clearly represented target labels and removed those images with irrelevant labels. For the Machine usage QA and Traditional management, annotators manually created categories based on textbooks and websites, and searched for corresponding images with redistribution licenses. For Weed identification, we use existing dataset [11, 23, 24, 42]. For this Weed identification category, we will provide a simple code to download the data and crop images and assign bounding boxes for Weed Identification, allowing users to work with the existing dataset without redistributing the images. Through these image selection processes, the selected images are high-quality 4,218 representative images. The detailed distribution of tasks and categories for each task is summarized in Figure 3.

QA Annotation. For the DMN, CMN, MQA, and TM tasks, all question-answer pairs were created manually, independent of any LLMs or VLMs. Annotators used GPT only for sentence rephrasing, but they were prohibited from using any knowledge from GPT for QA creation. Annotators were allowed to refer to textbooks, academic journals, and other authoritative sources in the field to ensure the accuracy and depth of the dataset. This took around 150 man-

Model	DID	DMN	PID	WID	CMN	MQA	TM	Overall (all)	Overall (subset)
Random Choice	21.77	15.64	20.40	17.90	16.06	22.11	19.31	19.03	19.11
Human	25.00	22.50	45.00	20.00	36.25	57.50	51.25	-	36.79
Closed-Source Vision Language Models (VLMs)									
GPT-4o mini [33]	53.60	80.67	60.04	35.14	64.23	70.96	69.80	62.06	69.65
GPT-4o [32]	64.18	87.35	77.76	44.17	75.43	82.84	82.43	73.45	79.26
Gemini1.5-Flash [46]	55.06	79.96	70.04	50.90	64.72	78.22	73.27	67.45	68.82
Gemini1.5-Pro [46]	62.92	81.55	74.45	55.17	71.05	82.84	77.72	72.24	69.74
Open-Source Vision Language Models (VLMs)									
EMU2Chat [44]	42.01	48.33	43.75	23.81	40.39	37.62	47.77	40.53	33.84
LLaVA-Next-8B [19]	45.47	72.58	43.01	30.05	54.26	56.11	57.46	51.28	57.84
LLaVA-Next-72B [19]	54.95	80.00	49.81	26.98	66.92	66.11	70.38	59.31	64.36
QwenVLM-7B [5]	51.26	80.49	63.97	33.17	66.42	76.24	77.48	64.15	66.41
QwenVLM-72B [5]	57.99	87.87	73.35	34.48	75.91	80.86	84.16	70.66	72.45
CogVLM-19B [50]	29.16	53.78	52.39	25.45	54.01	71.62	66.09	50.36	44.27
LLaVa-7B [18]	36.02	62.74	38.79	24.79	53.77	46.53	55.20	45.41	46.14
LLaVA-13B [18]	40.21	68.89	44.49	24.79	59.37	54.13	58.42	50.04	55.31

Table 2. **Results for the seven benchmark tasks with images.** We provide results for Random Choice, Human Validation, four closed-source VLMs, and open-source VLMs. Human validation was conducted by 28 people on a subset of 80 samples per task as a reference.

	DID	DMN	PID	WID	CMN	MQA	TM	Overall
GPT-4o [32]	1.93	72.58	18.75	1.00	40.39	25.08	48.27	29.71
LLaVA-Next-8B [19]	26.10	70.30	21.88	19.70	53.77	30.36	40.35	37.49

Table 3. **Results for the seven benchmark tasks with text only.** We additionally evaluate the model without image inputs, and the overall performance is close to random.

hours. We carefully annotate various types of questions. Examples of QAs that require expert knowledge are shown in Figure 4. All questions were carefully created so that image reference is necessary for answering.

Dataset Statistics. Following the dataset annotations and selections, our AgroBench comprises seven tasks with 4,342 QA pairs, as shown in Table 1. All tasks comprise a wide range of categories for detailed evaluation. Please refer to the supplementary materials for more details and examples.

4. Experiments

4.1. Experimental Settings

Baseline Models. We use four closed-source models: GPT-4o [32], GPT-4o mini [33], Gemini1.5-Pro [46], and Gemini1.5-Flash [46]. GPT-4o mini and Gemini1.5-Flash are down-scale versions of GPT-4o and Gemini1.5-Pro, respectively. We use eight open-source models: EMU2Chat [44], LLaVA-Next-8B [19], LLaVA-Next-72B [19], QwenVLM-7B [5], QwenVLM-72B [5], CogVLM-19B [50], LLaVa-7B [18], and LLaVa-13B [18]. For the details of these models, please refer to the supplementary material.

Human Results. We also present results from human participants for reference. We surveyed 28 students, each holding at least a bachelor's degree in agriculture, and asked them to answer 20 questions each. This created a test subset of 280 questions, with each question answered by two participants, resulting in a total of 560 responses. We averaged the results per task and reported the accuracy. Each participant was permitted to use a book or translator to look up word meanings but was prohibited from using the internet for searches. If participants were unsure of the answer, they were asked to provide the response they believed to be most accurate.

Evaluation Protocol. Importantly, our dataset evaluation is conducted per task, and overall scores are averaged based on the number of tasks, not the number of QAs. This prevents categories with a large number of evaluations from becoming dominant. We adopt an exact matching approach for our five-option questions. If the model's response matches the option's letter or the answer sentence, we consider it correct. If the model's answer does not match any option, including cases where there is no answer or mul-

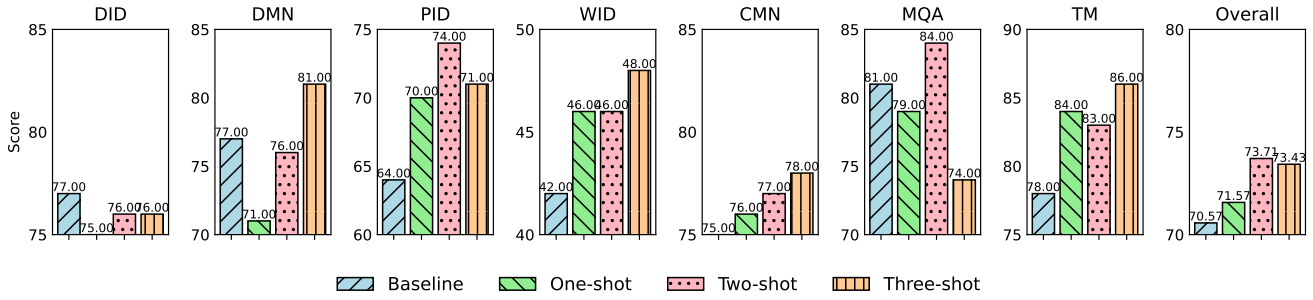

Figure 5. **Results of seven benchmark tasks with Chain of Thought (CoT).** Baseline indicates results without CoT. In the one-shot, two-shot, and three-shot settings, we provide one, two, and three CoT examples per task, respectively, to guide the model.

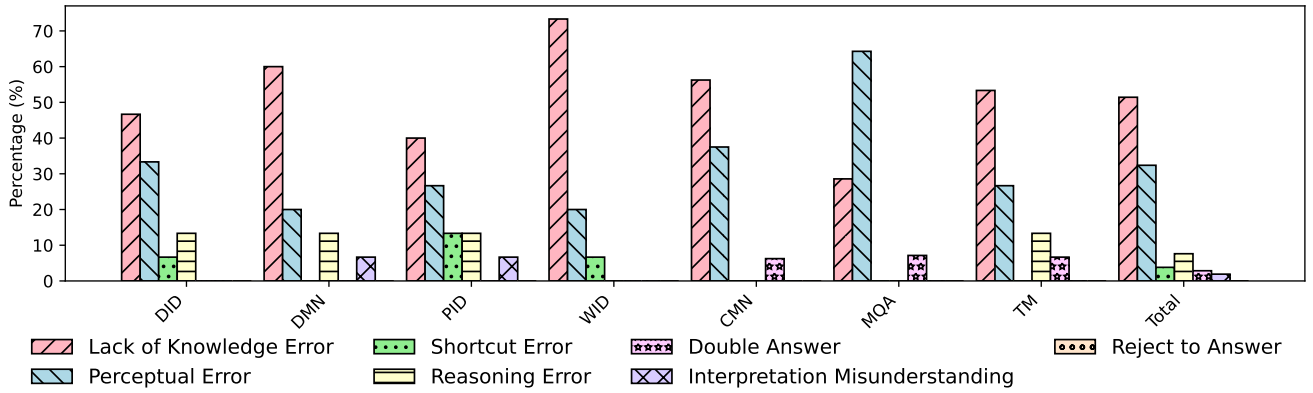

Figure 6. **Error analysis on seven benchmark tasks with GPT-4o.** We extract a maximum of 15 errors per task from the zero-shot CoT result. We manually analyze how they conclude the incorrect answer.

tiple answers, we consider it incorrect.

4.2. Main Results

Here, we discuss the main results of our AgroBench evaluation performance. We evaluate our AgroBench using eight open-source VLMs and four closed-source VLMs with APIs.

Challenges of the AgroBench. We show the main results in Table 2. The most difficult task is Weed Identification (WID), on which most open-source VLMs perform at around the random score. The highest WID accuracy, 55.17%, was achieved by Gemini 1.5-Pro. This suggests that the knowledge about weeds is not as fully trained as that of crops. All models' Disease Identification (DID) results are lower than their Disease Management (DMN) results. This means that VLM models can gain contextual information, but there is still room for perceptual improvement.

Model Comparison. Overall, closed-source VLMs achieve better results than open-source VLMs and achieve higher performance than humans. GPT-4o model achieved the highest score in overall performance. Among the open-source models, QwenVLM-72B achieves the best result in overall accuracy, which is comparable or even superior to open-source VLMs in some tasks. QwenVLM-72B achieves satisfactory results on both identification and question-answering benchmarks on AgroBench.

4.3. Ablations

Context Reliance. We further evaluate the results on AgroBench using text-only input to determine whether visual information is necessary to answer the questions. Table 3 presents the experimental results across seven benchmark tasks on AgroBench. With text-only input, performance drops across all models, confirming the reliance on visual information. However, for both models, the DMN, CMN, and TM tasks maintain significantly higher accuracy than random selection. Although we ensured that our questions do not include disease names or appearance-related traits, models tend to infer answers based on estimation. This suggests that many disease types share common management strategies, such as avoiding humidity, preferring cooler temperatures, or pruning infected parts, allowing models to predict the most likely option. A similar pattern is observed in the CMN and TM tasks, where models can make educated guesses based on contextual cues. (See the supplementary material for further detailed examples.)

Chain of Thought. We evaluate the effectiveness of Chain of Thought (CoT) reasoning. In the one-shot setting, we

(a) Lack of Knowledge

Q: What disease is affecting the plant?

GPT-4o: ..., **Sclerotinia blight**: Characterized by white, cottony mycelium and black sclerotia on stems. The image does not show these symptoms. *(not enough knowledge)* ..., the plant is likely affected by **D. Cylindrocladium black rot**. *(This is more common on root)*

A: Sclerotinia blight

(b) Perceptual Error

Q: What activity is being performed using the tractor-pulled machine shown in the image?

GPT-4o:The machine seems to be designed to lift plants from the soil, ..., and there are plants being lifted from the ground *(There are no harvested ctops)*.... the activity being performed is: D. Harvesting potato tubers from the soil using a tractor-pulled harvesting machine.

A: Cutting and chopping the stems and leaves using a tractor-pulled leaf chopper.

Figure 7. **Error examples of GPT-4o.** Examples illustrate the two main error types: Lack of Knowledge and Perceptual Error.

provide a single CoT reasoning example along with an image and the corresponding prompt for each task to guide the model in answering the question. We select 100 QAs per task for evaluation and use GPT-4 as the VLM. Figure 5 presents the accuracy of CoT reasoning with the baseline results. While CoT achieves slightly higher accuracy, the improvement is not significant compared to the baseline. However, CoT demonstrates effectiveness in certain tasks, such as PID, WID, CMN, and TM. For instance, in PID, the VLM performs step-by-step and careful reasoning (see the supplementary material for more detailed examples). In WID, the most challenging task, CoT provides useful examples that may help the model make predictions. Overall, CoT contributes to accuracy, but we observe performance saturation in the three-shot setting. For further detailed examples, please refer to the supplementary material.

4.4. Error Analysis

Here, we analyze the mistake types that occur depending on the agricultural tasks. We bring out up to 15 failure examples per task from the zero-shot CoT results and manually analyze how they reached the incorrect answers.

Lack of Knowledge (51.92%). This includes cases where VLM can't accurately describe the appearance or relevant knowledge of a choice (e.g., VLM fails to describe disease symptoms or insect characteristics) or lacks context (e.g., VLM doesn't know how to treat diseased crops or manage crops for high yield). Figure 7a shows an example of a Lack of Knowledge case. When the VLM analyzes the correct answer option for Sclerotinia blight, it fails to describe the symptoms of discoloration and wilting. Additionally, its incorrect answer choice is more commonly associated with the root rather than the stem. These errors are based on a Lack of Knowledge, suggesting VLMs need more detailed categories and domain-specific training.

Perceptual Error (32.69%). This indicates that VLM can't pay attention or recognize the answer-related part in the image (e.g., can't recognize the green insect on the leaf), and VLM misunderstands the image, leading to incorrect answers. Figure 7b shows an example of the model hallucinating and misunderstanding the situation. First, the VLM incorrectly identifies the machine as one used for lifting plants from the soil. Then, it describes the scene as if there are harvested crops, even though no harvested crops are present. These errors can be mitigated by enhancing the VLM's perception abilities for domain-specific classes. Additionally, improving general perception capabilities, including reducing hallucinations, can further contribute to VLM performance.

Reasoning Error (7.6%). Reasoning error involves the VLM can describe the options correctly, but can't compare them step by step and conclude the wrong answer. This error is relatively low compared to the existing work [55] since AgroBench requires more specific knowledge and doesn't include reasoning relying on problems (e.g., math).

Other Errors (7.79%). For the other errors, we observe Shortcut Error (The VLM can pick up the two candidate options correctly but conclude the answer without comparing the candidates), Double Answer Error (Concluding two answers are correct), Interpretation Misunderstanding (VLM misleading the question and conclude wrong answer), and Reject to Answer (VLM conclude there is no answer).

5. Conclusion

In this paper, we develop AgroBench, a comprehensive benchmark dataset for VLMs in the agricultural domain, covering a state-of-the-art range of categories. Our dataset comprises seven benchmark tasks encompassing key research areas in agricultural engineering as well as real-world challenges faced by farmers. AgroBench contributes to agricultural VLM research by addressing the lack of datasets for comprehensive multi-modal model evaluation, validated by human experts.

In our evaluation, VLMs exhibit strengths across different tasks. However, in several tasks such as weed identification and disease identification, all models show room for improvement. Our error analysis reveals that most failures are due to a lack of knowledge (51.92%), suggesting that VLMs require more specialized agricultural knowledge. Our dataset will facilitate agricultural VLM research, enabling broad category and task evaluation to support sustainable, automated agriculture.

6. Acknowledgment

This work was supported by the AIST KAKUSEI Project (FY2024) and JST FOREST Grant Number JPMJFR206F. We would like to thank the Agricultural Administration Division, Department of Agriculture, Hokkaido Government, for providing some of the images. We are also grateful to Daniel Steininger for contributing images from the CropAndWeed dataset to our dataset. We used ABCI 3.0 provided by AIST and AIST Solutions.

References

[1] Petchiammal A, Briskline Kiruba S, Murugan D, and Pandarasamy Arjunan. Paddy doctor: A visual image dataset for automated paddy disease classification and benchmarking. *IEEE Dataport*, 2022. 2, 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 23716–23736, 2022. 3

[3] Abderraouf Amrani, Ferdous Sohel, Dean Diepeveen, David Murray, and Michael G.K. Jones. Deep learning-based detection of aphid colonies on plants from a reconstructed brassica image dataset. *Computers and Electronics in Agriculture*, 205:107587, 2023. 2

[4] Muhammad Awais, Ali Husain Salem Abdulla Alharthi, Amandeep Kumar, Hisham Cholakkal, and Rao Muhammad Anwer. Agrogpt: Efficient agricultural vision-language model with expert tuning. *arXiv preprint arXiv:2410.08405*, 2024. 2

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3, 6

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 3

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 2

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Alexander Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. 2023. 3

[9] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Tvbench: Redesigning video-language evaluation. 2024. 3

[10] Mamta Gehlot, Rakesh Saxena, and Geeta Gandhi. "tomato-village": a dataset for end-to-end tomato disease detection in a real-world environment. *Multimedia Systems*, 29:1–24, 2023. 2, 3

[11] Benedikt Geisler. Perrenial plants detection, 2021. 5

[12] Sebastian Haug and Jörn Ostermann. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *Proc. European Conference on Computer Vision (ECCV)*, pages 105–116. Springer, 2014. 3

[13] Atuhurra Jesse, N'guessan Yves-Roland Douha, and Pabitra Lenka. Image classification for cssvd detection in cacao plants. *arXiv preprint 2405.04535*, 2024. 2, 3

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 3

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 3

[16] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024. 3

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023. 2, 3

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 3, 6

[19] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 6

[20] Xiang Liu, Zhaoxiang Liu, Huan Hu, Zezhou Chen, Kohou Wang, Kai Wang, and Shiguo Lian. A multimodal benchmark dataset and model for crop disease diagnosis. In *Proc. European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[21] Xinpeng Liu, Kanyu Xu, Risa Shinoda, Hiroaki Santo, and Fumio Okura. Masks-to-skeleton: Multi-view mask-based

tree skeleton extraction with 3d gaussian splatting. *Sensors*, 25(14), 2025. 2

[22] Shenglian Lu, Wenkang Chen, Xin Zhang, and Manoj Karkee. Canopy-attention-yolov4-based immature/mature apple fruit detection on dense-foliage tree architectures for early crop load estimation. *Computers and Electronics in Agriculture*, 193:106696, 2022. 2

[23] Yuzhen Lu. Cottonweeddet3. kaggle, 2022. 5

[24] Simon Leminen Madsen, Solvejg Kopp Mathiassen, Mads Dyrmann, Morten Stigaard Laursen, Laura-Carlota Paz, and Rasmus Nyholm Jørgensen. Open Plant Phenotype Database of Common Weeds in Denmark, 2020. 5

[25] Abdul Khalique Maitlo, Abdul Aziz, Hassnian Raza, and Neelam Abbas. A novel dataset of guava fruit for grading and classification. *Data in Brief*, 49:109462, 2023. 2

[26] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. 2, 3

[27] João Matos, Shan Chen, Siena Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David Restrepo, Luis F. Nakayama, Jose M. M. Pascual-Leone, Guergana Savova, Hugo Aerts, Leo A. Celi, A. Ian Wong, Danielle S. Bitterman, and Jack Gallifant. Worldmedqa-v: a multilingual, multimodal medical examination dataset for multimodal language models evaluation, 2024. 3

[28] Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *WACV*, pages 1516–1525, 2020. 2, 3

[29] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 2016. 3

[30] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models, 2023. 3

[31] Alex Olsen, Dmitry A Konovalov, Bronson Philippa, Peter Ridd, Jake C Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, et al. Deepweeds: A multiclass weed species image dataset for deep learning. *Scientific reports*, 9(1):2058, 2019. 3

[32] OpenAI. Gpt-4o, 2024. Accessed: 2024-11-13. 2, 3, 6

[33] OpenAI. Gpt-4o mini, 2024. Accessed: 2024-11-13. 2, 6

[34] Chiranjit Pal, Imon Mukherjee, Sanjay Chatterji, Sanjoy Pratihar, Pabitra Mitra, and Partha Pratim Chakrabarti. Indian rice disease dataset (irdd), 2023. 2, 3

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[36] Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models, 2024. 3

[37] Rahman Sanya, Ann Lisa Nabiryo, Jeremy Francis Tusubira, Sudi Murindanyi, Andrew Katumba, and Joyce Nakatumba-Nabende. Coffee and cashew nut dataset: A dataset for detection, classification, and yield estimation for machine learning applications. *Data in Brief*, 52:109952, 2024. 2

[38] Nabila Husna Shabrina, Siwi Indarti, Rina Maharani, Dinar Ajeng Kristiyanti, Irmawati, Niki Prastomo, and Tika Adilah M. A novel dataset of potato leaf disease in uncontrolled environment. *Data in Brief*, 52:109955, 2024. 2

[39] Risa Shinoda, Hirokatsu Kataoka, Kensho Hara, and Ryozo Noguchi. Transformer-based ripeness segmentation for tomatoes. *Smart Agricultural Technology*, 4:100196, 2023. 2

[40] Risa Shinoda, Kuniaki Saito, Shohei Tanaka, Tosho Hirasawa, and Yoshitaka Ushiku. Sbs figures: Pre-training figure qa from stage-by-stage synthesized images, 2024. 3

[41] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. Plantdoc: A dataset for visual plant disease detection. In *Proc. ACM IKDD CoDS and COMAD*, page 249–253, 2020. 2, 3

[42] Daniel Steininger, Andreas Trondl, Gerardus Croonen, Julia Simon, and Verena Widhalm. The cropandweed dataset: A multi-modal learning approach for efficient crop and weed manipulation. In *WACV*, pages 3729–3738, 2023. 3, 5

[43] Daniel Steininger, Julia Simon, Andreas Trondl, and Markus Murschitz. Timbervision: A multi-task dataset and framework for log-component segmentation and tracking in autonomous forestry operations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 2

[44] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023. 3, 6

[45] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 2, 3

[46] Google Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 2, 3, 6

[47] Sandip Thite, Yogesh Suryawanshi, Kailas Patil, and Prawit Chumchu. Sugarcane leaf dataset: A dataset for disease detection and classification for machine learning applications. *Data in Brief*, 53:110268, 2024. 2, 3

[48] Liqiong Wang, Teng Jin, Jinyu Yang, Ales Leonardis, Fangyi Wang, and Feng Zheng. Agri-llava: Knowledge-infused large multimodal assistant on agricultural pests and diseases. *arXiv preprint arXiv:2412.02158*, 2024. 2

[49] Rujing Wang, Liu Liu, Chengjun Xie, Po Yang, Rui Li, and Man Zhou. Agripest: A large-scale domain-specific benchmark dataset for practical agricultural pest detection in the wild. *Sensors*, 21(5), 2021. 2, 3

[50] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming

Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. In *NeurIPS*, 2024. 2, 3, 6

[51] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3

[52] Tianqi Wei, Zhi Chen, Zi Huang, and Xin Yu. Benchmarking in-the-wild multimodal disease recognition and a versatile baseline. In *Proc. ACM International Conference on Multimedia (ACMMM)*, 2024. 2, 3

[53] Xiaoping Wu, Chi Zhan, Yukun Lai, Ming-Ming Cheng, and Jufeng Yang. Ip102: A large-scale benchmark dataset for insect pest recognition. In *CVPR*, pages 8787–8796, 2019. 2, 3

[54] Momchil Yordanov, Raphaël d'Andrimont, Laura Martinez-Sanchez, Guido Lemoine, Dominique Fasbender, and Marijn van der Velde. Crop identification using deep learning on lucas crop cover photos. *Sensors*, 23(14), 2023. 2

[55] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567, 2024. 2, 3, 8

[56] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 2, 3