🌲 AgriGPT-VL: An Unified Suite for Agricultural Vision–Language Understanding

Bo Yang
Zhejiang University
boyang30@zju.edu.cn

Lanfei Feng
Zhejiang University
22451116@zju.edu.cn

Yunkui Chen
Zhejiang University
22351048@zju.edu.cn

Yu Zhang
Zhejiang University
22421173@zju.edu.cn

Xiao Xu
Zhejiang University
3200105334@zju.edu.cn

Jianyu Zhang
Zhejiang University
jianyu.zhang@zju.edu.cn

Nueraili Aierken
Zhejiang University
nureli@zju.edu.cn

Runhe Huang
Hosei University
rhuang@hosei.ac.jp

Hongjian Lin
Zhejiang University
linhongjian@zju.edu.cn

Yibin Ying
Zhejiang University
ibeying@zju.edu.cn

Shijian Li*
Zhejiang University
shijianli@zju.edu.cn

Abstract

*Despite rapid advances in multimodal large language models, agricultural applications remain constrained by the scarcity of domain-tailored models, curated vision–language corpora, and rigorous evaluation. To address these challenges, we present the **AgriGPT-VL Suite**, a unified multimodal framework for agriculture. Our contributions are threefold. First, we introduce **Agri-3M-VL**, the largest vision–language corpus for agriculture to our knowledge, curated by a scalable multi-agent data generator; it comprises 1M image–caption pairs, 2M VQA (Visual Question Answering) pairs, 50K expert-level VQA, and 15K GRPO reinforcement learning dataset. Second, we develop **AgriGPT-VL**, an agriculture-specialized vision–language model trained via a progressive curriculum of textual grounding, multimodal shallow/deep alignment, and GRPO refinement. This method achieves strong multimodal reasoning while preserving text-only capability. Third, we establish **AgriBench-VL-4K**, a compact yet challenging evaluation suite with a multi-metric evaluation and an LLM-as-a-judge framework. Experiments show that AgriGPT-VL outperforms leading general-purpose VLMs on AgriBench-VL-4K, achieving higher pairwise win rates in the LLM-as-a-judge evaluation. Meanwhile, it remains competitive on the text-only AgriBench-13K with no noticeable degradation of language ability. Ablation studies further confirm consistent gains from our alignment and*

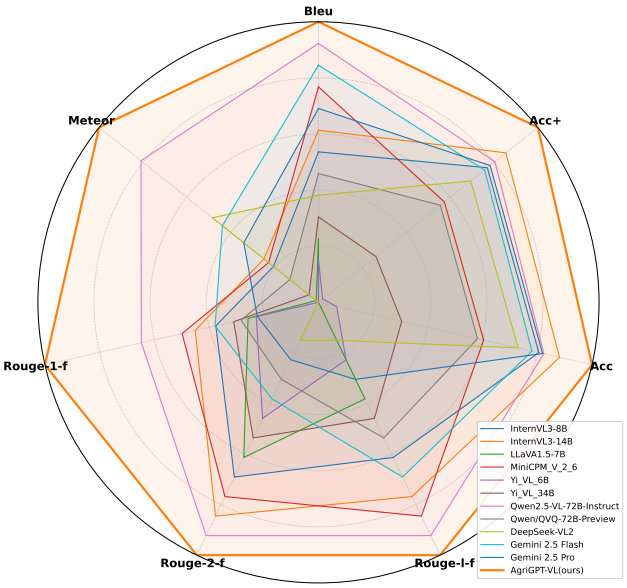

Figure 1. AgriGPT-VL achieves leading performance on AgriBench-VL-4K.

GRPO refinement stages.

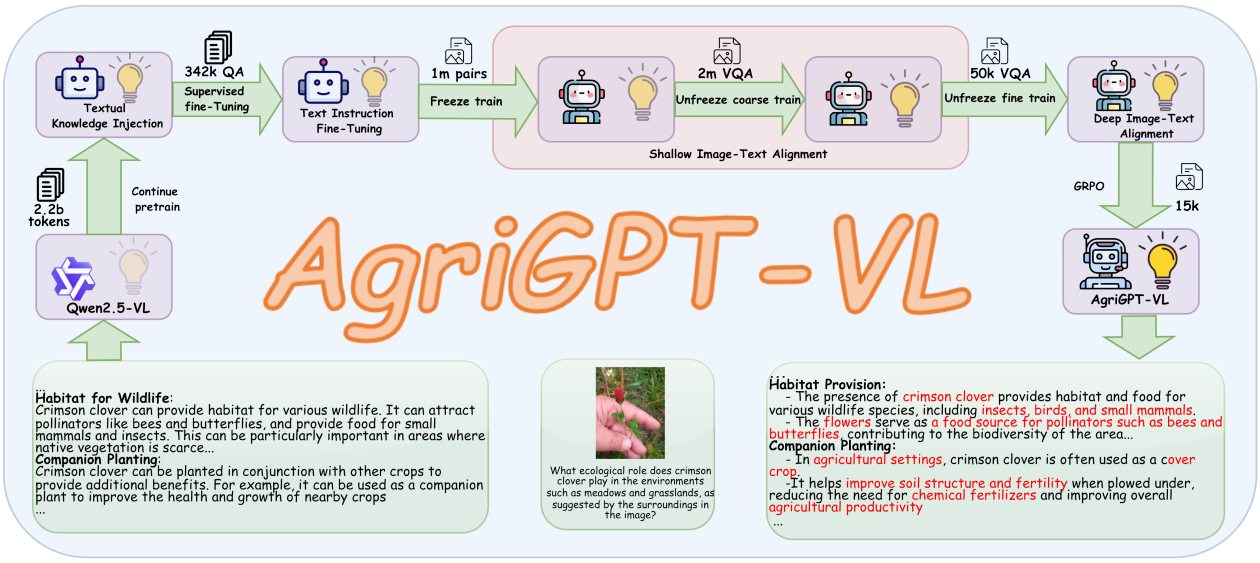

Figure 2. Overview of the AgriGPT-VL training pipeline, showing the progression from textual tuning to shallow and deep image–text alignment, followed by GRPO refinement, together with the datasets used at each stage.

1. Introduction

The convergence of AI with critical sectors like agriculture presents a significant opportunity to address global challenges such as food security and sustainable resource management [12, 19, 46]. With the increasing challenges posed by climate change, resource scarcity, and population growth, intelligent agricultural decision-making is becoming indispensable [18, 21, 42]. In recent years, multimodal large language models (MLLMs) have demonstrated remarkable progress in integrating vision and language, enabling tasks such as captioning, visual question answering (VQA), and multimodal reasoning [4, 10, 56]. While Multimodal Large Language Models (MLLMs) excel at integrating vision and language on general web data [17, 44], they are ill-equipped for the agricultural domain. The knowledge required for tasks in crop and soil science is highly specialized and absent from standard pre-training corpora [28, 50]. Consequently, existing MLLMs struggle with agricultural terminology, exhibit factual inaccuracies, and fail to provide reliable, context-aware support for real-world farming operations [41, 51, 54].

Several attempts have been made to build agricultural language models, such as AgriBERT [41], AgriLLM [16], AgroLLM [43], and AgroGPT [6]. These efforts show the value of domain-specific adaptation but are often constrained to text-only settings or narrow task coverage. AgriGPT [54], introduced the first agriculture-specialized LLM ecosystem with a curated instruction dataset(Agri-342K), a retrieval-enhanced reasoning module (Tri-RAG), and a benchmark suite (AgriBench-13K). While effective for textual reasoning, AgriGPT lacked visual reasoning ability and thus could not address multimodal agricultural tasks such as pest recognition or crop diagnosis. On the other hand, general-purpose MLLMs such as InternVL [11], Qwen-VL [47], Gemini [4], and LLaVA [32] demonstrate strong vision–language capabilities but are trained primarily on internet-scale data describing common objects, scenes, and events, which fail to capture agricultural semantics. As a result, these models suffer from hallucinations, poor transferability, and lack of reasoning ability in agriculture-specific scenarios. Related domains such as medicine developed specialized multimodal LLMs, highlighting the need for a comparable ecosystem in agriculture.

Our contributions can be summarized as follows:

- **Agri-3M-VL Dataset & Data Generator.** We build a transferable, reusable multi-agent Data Generator and use it to curate **Agri-3M-VL**: **1M** image–caption pairs, **2M** high-quality VQA pairs, **50K** expert-level VQA, and **15k** rewarded GRPO reinforcement learning dataset.
  To the best of our knowledge, this is the largest agriculture vision–language corpus to date.
- **AgriGPT-VL & Curriculum Training.** Using a progressive curriculum, we train the agriculture-specialized VL model **AgriGPT-VL**, as shown in figure 1, which surpasses most flagship models in capability.
- **AgriBench-VL-4K & Evaluation Framework.** We construct a comprehensive and challenging benchmark with **2,018** open-ended VQA and **1,858** image-grounded single-choice questions (two per image for cross-consistency).
- **Open Resources.** All resources will be released as open-source to ensure full reproducibility and enable deploy-

ment in low-resource agricultural scenarios.

2. Related Work

2.1. Text-Only Language Models in Agriculture

Pioneering work in agricultural AI largely focused on the language modality. Early models such as AgriBERT [41] adapted language model pre-training to domain-specific text corpora. Subsequent efforts, including AgriLLM [16], AgroLLM [43], and AgriGPT [54], advanced this paradigm by developing large-scale instruction datasets like Agri-342K and text-only benchmarks such as AgriBench-13K [54]. For instance, Zhu et al. [62] reviewed the progression of text-only and multimodal agricultural LLMs, highlighting the transition from domain adaptation to instruction-based fine-tuning. Moreover, Yu and Lin [58] proposed a framework leveraging LLMs for agricultural knowledge inference and consultation, suggesting broader utility beyond QA. While these models demonstrated strong textual understanding, their primary limitation was the absence of visual reasoning ability, restricting their applicability to tasks that do not require visual interpretation.

2.2. Emergence of Multimodal Agricultural Systems

The integration of visual data marked a critical evolution in agricultural AI. Foundational datasets like PlantVillage [26] and IP102 [52] provided large-scale image collections for specific recognition tasks, such as pest and disease identification. More recent works have begun to build multimodal models and benchmarks with broader capabilities. For instance, Agri-LLaVA [48], AgriCLIP [37], and LLMI-CDP [49] introduced vision–language abilities, while datasets like VL-PAW [57] and benchmarks like AgMMU [20], AgroBench [45], and AgriEval [53] introduced tasks such as VQA and captioning. Other studies such as Zhu et al. [62] provide a systematic review of the current landscape, while Yu and Lin [58] and Arshad et al. [5] explore concrete frameworks or empirical evaluations of VLMs in agricultural use cases. However, these multimodal resources often remain limited in scale, are restricted to narrow recognition tasks, or lack rigorous, large-scale quality control, representing disparate efforts rather than a cohesive foundation.

2.3. The Need for a Unified Vision–Language Ecosystem

The limitations of prior work highlight a clear need for a comprehensive and unified framework. While previous efforts have made valuable contributions to datasets, models, or benchmarks individually, progress has been hampered by the lack of a single ecosystem that integrates all three components at scale. To address this fragmentation, our work introduces a cohesive suite of resources. Our **Agri-3M-VL dataset** provides scale and quality; our **AgriGPT-VL** model handles complex reasoning beyond simple recognition; and our **AgriBench-VL-4K** benchmark enables robust, multifaceted evaluation. Together, these components form the kind of unified foundation we argue is necessary for the next generation of agricultural AI.

3. AgriGPT-VL

3.1. Agri-3M-VL Dataset

Constructing training data is a fundamental challenge in developing multimodal large language models. To address this, we introduce the Data Generator, a transferable paradigm for systematically transforming raw images into high-quality multimodal instructions. The generator is designed not only for agriculture but also as a generalizable methodology that can be applied to other scientific domains where multimodal resources remain scarce or noisy.

As shown in Figure 4, we aggregated a wide range of datasets covering pests and diseases, insects, crops, weeds, and fruits. Specifically, the PlantVillage dataset contains 54,305 images across 38 classes [3]. For insect-related data, we included 6,878 images covering 166 fine-grained insect species from the Species196 dataset [24], and the Insect Foundation dataset with 317,128 images spanning 38,867 fine-grained insect classes [59], totaling 324,006 images and 39,033 classes. In the crop and weed domain, the SelectDataset provides 558,930 images over 2,958 categories [14]. For fruits, we incorporated Fruits-360 with 97,255 images and 206 categories [36], and Fresh-Rotten Fruit with 30,357 images and 18 categories [15], amounting to a combined 157,969 images and 224 classes. Altogether, these datasets cover **1,064,853** images and **42,253** fine-grained categories, nearly encompassing the full agricultural visual landscape.

However, these raw datasets suffer from several limitations: many lack descriptive annotations, exhibit inconsistent labeling, and cannot be directly used for multimodal model training. These shortcomings necessitate our proposed Data Generator, which systematically transforms such raw images into structured, instruction-ready corpora. Through several stages of processing, the Data Generator enables the creation of a large-scale, high-quality multimodal training corpus suitable for agricultural vision–language modeling.

As shown in figure 3, the Data Generator transforms multi-source agricultural images into instruction-ready corpora via four stages—caption generation, instruction synthesis, multi-agent refinement and instruction filtering yielding 1M image captions, 2M high-quality VQA, a 50K expert-level VQA, and 15k GRPO reinforcement learning dataset.The detailed high-quality VQA are illustrated in figure 5.

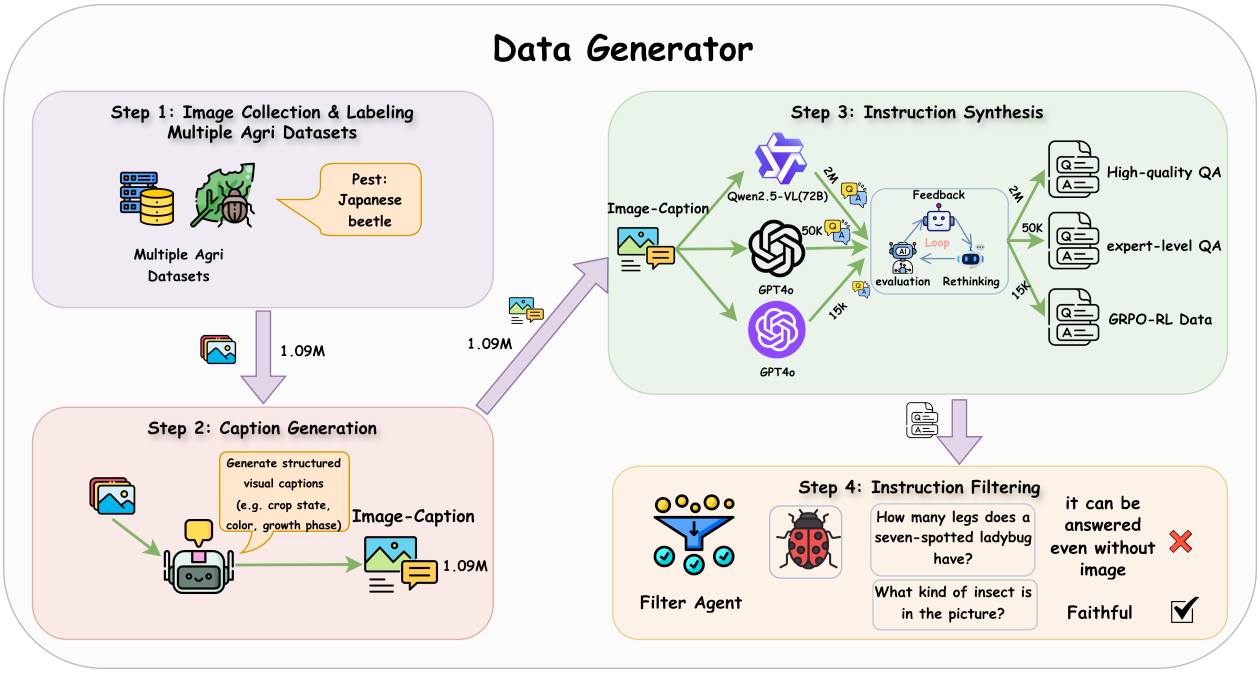

Figure 3. Data Generator: A multimodal instruction data generation pipeline.

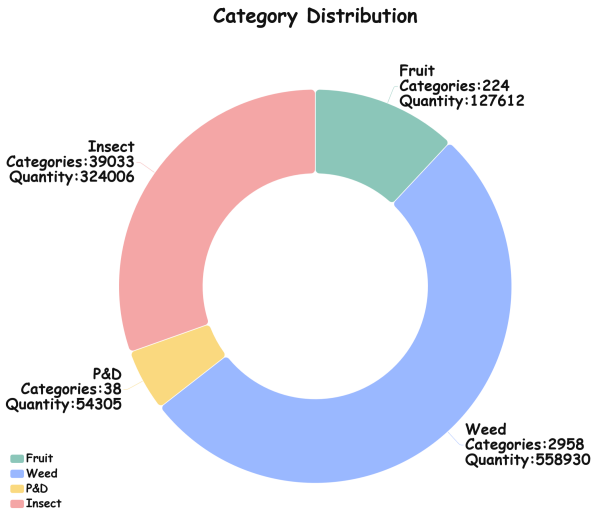

Figure 4. Category distribution of the dataset.

(1) Caption Generation. For the collected images spanning pests and diseases, insects, weeds, and fruits, we first generate structured visual captions. These captions describe observable attributes such as crop growth stage, leaf color, fruit maturity, or pest morphology. For example, an image of diseased tomato leaves is captioned with information about lesion color and spread, while a fruit image records ripeness stage and external texture. In total, this stage yields about 1 million image–caption pairs, providing a descriptive foundation for subsequent instruction synthesis. The significance of this step is that captions transform raw visual data into semantically rich text, enabling downstream models to link domain-specific imagery with meaningful language.

(2) Instruction Synthesis. Building upon the image–caption pairs, we employ large vision–language models (e.g., Qwen2.5-VL 72B, GPT-4o) to generate diverse instructions and answers. This stage produces multiple types of VQA: high-quality factual queries, expert-level reasoning tasks, and interactive multimodal dialogues. For instance, a weed image may lead to questions such as "What species of weed is shown?" (recognition) or "What is the likely impact of this weed on crop yield?" (reasoning). Altogether, we synthesize approximately 2 million VQA samples, covering both open-ended and single-choice formats. This step is essential because it elevates the dataset from simple recognition to instruction-following reasoning, directly aligning with the needs of multimodal LLMs.

(3) **Multi-Agent Refinement.** To further construct high-quality VQA data on top of the Image–Caption corpus, we design a protocol-guided multi-agent refinement architecture and adopt Qwen2.5-72B as the core execution model to balance efficiency and computational cost. The architecture consists of three expert agents—**Feedback**, **Evaluation**, and **Rethinking**—which collaborate through a structured "generate → assess → revise" loop. The **Feedback agent** generates an initial question–answer draft based on the image and its caption, with protocol constraints requiring that the content reference only visible entities and attributes,

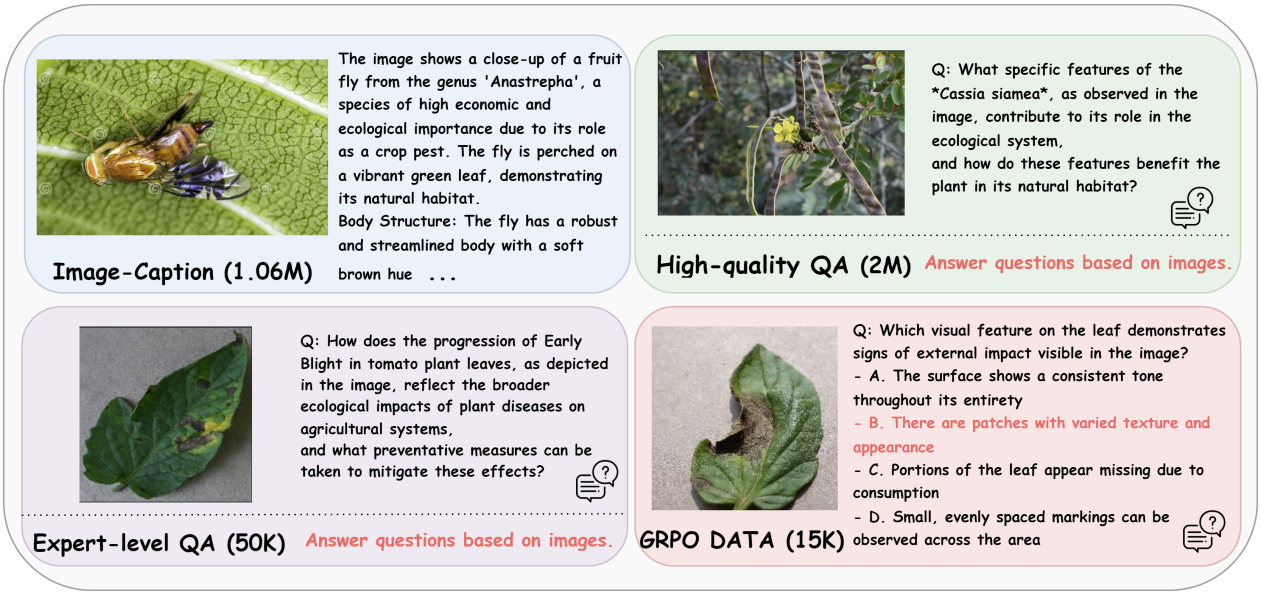

Figure 5. The four types of hierarchical training data constructed for AgriGPT-VL.

avoid out-of-image inferences, and maintain clear visual evidence. The **Evaluation agent** agent performs structured quality assessment along correctness, clarity, and completeness (each scored within $[0, 1]$) and also produces concise diagnostic feedback identifying inconsistencies, ambiguity, or insufficient grounding; any score below the unified threshold of 0.85 triggers revision. The **Rethinking agent** revises the question and answer according to this feedback and performs a self-consistency check by independently generating the answer three times, requiring an agreement of at least 0.85; otherwise, further revision is initiated. This protocolized refinement loop typically converges in approximately 2.3 rounds, and we retain only samples that satisfy all quality requirements, yielding about 2M high-quality VQA pairs. We further select 50K samples for multi-round verification and polishing using GPT-4o to construct a high-quality supervised fine-tuning subset, and additionally create a 15K GRPO preference dataset, also with GPT-4o, for reinforcement learning. Empirically, about 8% of initial drafts are filtered due to insufficient correctness or grounding, and manual inspection of a 5K subset confirms that all three quality dimensions consistently exceed 0.86, demonstrating the effectiveness of the multi-agent refinement architecture in reducing hallucination and improving data reliability.

(4) Instruction Filtering. Finally, we introduce a **Filter agent** to automatically identify and discard instructions that are irrelevant to the image or potentially hallucinated. The agent evaluates each instruction along three protocol-defined dimensions—correctness, image-dependence, and grounding validity—and marks a sample for removal whenever any dimension falls below the required threshold. As a result, generic questions unrelated to the image (e.g., "How many legs does a seven-spotted ladybug have?") are correctly filtered out, while questions that genuinely rely on visual evidence (e.g., "What kind of insect is in the picture?") are retained. Quantitatively, about 0.6% of the initial instructions are accurately identified as irrelevant or factually inconsistent; in a human audit of 300 randomly sampled instructions, none of the filtered data exhibited such issues. This filtering step further strengthens factual alignment, suppresses hallucination propagation, and improves the overall trustworthiness of the training data.

Each stage is complementary: caption generation provides semantic grounding, instruction synthesis injects reasoning diversity, multi-agent refinement structures feedback-driven selection, and instruction filtering enforces factual reliability. Together, they form a robust agricultural multimodal dataset that not only supports AgriGPT-VL training but also serves as a blueprint for dataset construction in other scientific domains.

3.2. AgriGPT-VL Model Training

As shown in figure 2. This section details our training paradigm for AgriGPT-VL, which follows a progressive curriculum: textual grounding first, then vision–language alignment. We first consolidate domain knowledge and instruction style on text-only data, and then align vision and language on synthesized multimodal supervision with an easy-to-hard schedule.

Stage-1 (Text-only). Starting from Qwen2.5-VL, we conduct continual pretraining on about 200K documents ($\approx$

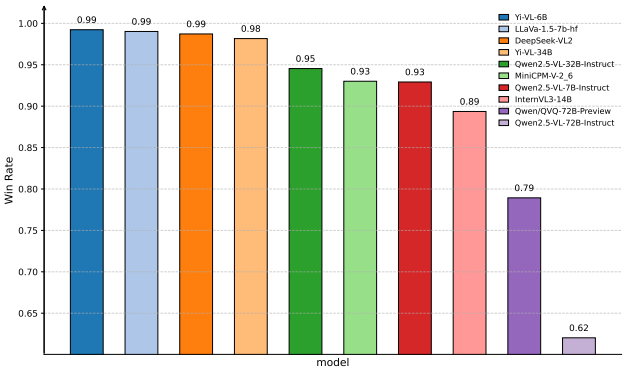

Figure 6. Pairwise win rate of vision-language models vs. AgriGPT-VL (Qwen2.5-VL-72B-as-a-judger)

2.2B tokens) to inject agricultural terminology and background knowledge, followed by supervised instruction tuning on Agri-342K [54]. A held-out split of AgriBench-13K [54] is used for early stopping and calibration prior to multimodal alignment.

Stage-2 (Curricular Alignment on Synthesized Multimodal Data). We adopt a three-step easy-to-hard sequence built on caption and VQA supervision, then preference optimization:

(2a) Shallow Alignment. We start with 1M image–caption pairs, keeping both the vision encoder and LLM component fully frozen. Only the connector and adapter layers are trained. Captioning tasks help establish a stable semantic bridge between vision and language modalities.

(2b) Deep Alignment (From Coarse to Full Training). Next, we train on 2M image–QA samples (two questions per image for cross-validation) and 50K GPT-4o-polished samples for supervised fine-tuning, covering recognition, attributes, diagnosis, and basic multi-hop reasoning. Using LoRA, we gradually unfreeze the vision encoder and LLM, enabling transition to full multimodal reasoning.

(2c) GRPO Optimization. We build 15K GRPO samples for reinforcement learning. GRPO rewards image-text consistency, internal logic, and verifiable terminology. Details are in Appendix A.2 and A.6.

3.3. AgriBench-VL-4K

Construction. To ensure the objectivity and reliability of evaluation, we construct AGRIBENCH-VL-4K with an emphasis on data independence, procedural transparency, and rigorous quality control. The benchmark consists of two components: 2,018 open-ended question–answer pairs and 1,858 single-choice questions. All items are regenerated from held-out images that are never used during training, preventing any overlap with the distributional patterns of the training data generator. Open-ended questions are deeply rewritten from structured captions, covering recognition, symptom/mechanism analysis, management recommendations, and simple multi-step reasoning; answers are normalized for terminology, synonyms, and measurement units. For the single-choice portion, two complementary questions are designed per image, and cross-question consistency is used to reduce random guessing. Distractors are mined from confusable taxa and frequently co-occurring agricultural conditions to increase discriminability and reduce generator-style bias.

Quality control and de-duplication. To further mitigate potential data contamination, we perform strict de-duplication at both the image and text levels: perceptual hashing and visual-feature similarity are used to remove near-duplicate images, and lexical as well as embedding similarity are applied to eliminate near-duplicate question–answer strings. De-duplication is conducted both across train–evaluation splits and within the evaluation split itself, ensuring that AGRIBENCH-VL-4K contains no residual patterns originating from the training set or the data generator. All remaining items then undergo manual review by annotators with backgrounds in agriculture and computer science, who jointly verify factual correctness, image-grounded evidence, and ambiguity resolution. In practice, approximately 7.8% of initial items (4% for open-ended QA and 11.5% for single-choice QA) are rejected due to factual inconsistencies or lack of image relevance.

4. Results

4.1. Comparative Experiment

We focus on two questions: (i) after progressively injecting domain knowledge and vision–language alignment, is textual competence preserved and strengthened; and (ii) in real image–language settings, does the model exhibit stronger visual grounding and agronomic reasoning—i.e., can it both choose correctly (discriminative robustness) and articulate evidence–based answers (generation quality). To this end, we evaluate text–only capability on AgriBench-13K [54]and multimodal capability on AgriBench-VL-4K. For discriminative evaluation, we report *Acc* (single–choice accuracy, scored per question) and Acc^{+} (image–level cross–consistency: both single–choice questions for the same image must be correct). For generation, we report BLEU [39], METEOR [9], and ROUGE-L [30] to measure terminology conformity, semantic coverage, and structural completeness.

We compare *AgriGPT-VL* against twelve representative vision–language models: InternVL-3-8B/3-14B [63], LLaVA-1.5-7B [31], MiniCPM-V-2.6 [38, 55], Yi-VL-6B and Yi-VL-34B [1, 2], Qwen-VL-7B [7], Qwen2.5-VL-72B-Instruct [8], Qwen-QVQ [40], DeepSeek-VL-1.2 [34], and Gemini-2.5 Flash/Pro [22, 23].

Table 1. Language-only evaluation (text capability only). Comparison of AgriGPT-VL with general VLMs on text tasks without images. **Bold** and underlined denote best and second-best per column.

Model	BLEU	Meteor	Rouge-1-f	Rouge-2-f	Rouge-L-f
InternVL-3-8B	5.52	23.07	24.14	5.69	23.08
InternVL-3-14B	8.53	27.56	26.75	**6.46**	25.56
LLaVA-1.5-7B	1.44	13.62	21.67	4.88	20.60
MiniCPM-V-2.6-8B	1.15	12.50	21.41	5.25	20.16
Yi-VL-6B	1.03	12.38	20.93	4.36	20.09
Yi-VL-34B	1.69	14.27	21.82	4.74	20.27
Qwen-VL-7B	7.70	30.17	24.16	4.97	22.86
Qwen2.5-VL-72B-Instruct	6.52	30.27	25.84	5.95	24.43
Qwen/QVQ-72B-Preview	2.48	17.31	17.54	3.63	16.75
DeepSeek-VL-1.2	6.37	29.67	22.10	4.93	20.93
Gemini-2.5-Flash	6.12	27.49	24.85	5.59	23.73
Gemini-2.5-Pro	4.34	23.69	24.16	4.45	22.30
AgriGPT-VL (ours)	**10.84**	**32.53**	**27.73**	6.36	**26.36**

Table 2. Performance comparison on the AgriBench-VL-4K benchmark. **Acc** and **Acc**$^+$ correspond to accuracy on single-choice visual reasoning questions, whereas the other metrics evaluate open-ended question–answering quality.

Model	Acc	Acc$^+$	BLEU	Meteor	Rouge-1-f	Rouge-2-f	Rouge-L-f
InternVL-3-8B	81.27%	66.31%	5.38	21.85	33.43	12.22	30.69
InternVL-3-14B	83.05%	69.21%	6.32	23.26	35.01	13.44	32.11
LLaVA-1.5-7B	62.33%	40.04%	2.39	15.55	30.96	11.09	28.57
MiniCPM-V-2.6-8B	76.53%	59.63%	7.12	22.57	36.02	13.39	33.36
Yi-VL-6B	63.89%	40.69%	2.21	15.21	30.32	10.59	27.88
Yi-VL-34B	69.48%	48.98%	2.83	16.61	32.05	10.98	29.42
Qwen2.5-VL-72B-Instruct	81.70%	67.49%	15.41	41.38	39.14	13.25	36.63
Qwen/QVQ-72B-Preview	76.00%	58.99%	4.55	19.43	31.46	9.09	29.53
DeepSeek-VL2	79.49%	63.72%	2.88	30.86	25.57	7.21	22.97
Gemini-2.5-Flash	80.68%	65.88%	9.12	29.38	33.47	10.00	31.33
Gemini-2.5-Pro	81.59%	66.74%	6.55	26.22	30.25	8.65	28.01
AgriGPT-VL (ours)	**85.84%**	**74.17%**	**26.27**	**47.55**	**46.52**	**20.09**	**43.81**

As shown in tables 1, on AgriBench-13K [54], AgriGPT-VL leads across mainstream text metrics, indicating that the progressive training does not sacrifice language ability; instead, it strengthens standardized use of agricultural terminology and canonical answer style, consolidating textual representations and providing a stable linguistic base for the subsequent multimodal stage.

As shown in table 2, on AGRIBENCH-VL-4K, we obtain the best results on all metrics, surpassing several flagship large models. Gains in *Acc* reflect more precise image–option matching; gains in *Acc*$^+$ demonstrate consistent semantics per image and stronger resistance to hard distractors (confusable taxa and co–occurring conditions), thereby mitigating chance guessing and better reflecting true capability. Improvements in Bleu [39], Meteor [9], and Rouge-1-f/Rouge-2-f/Rouge-L-f [30] further indicate three strengthened abilities: (1) visual evidence grounding and factor extraction (organs, colors/lesions, phenology); (2) agronomic multi-step reasoning (from symptoms to plausible causes and management consistent with scene constraints); and (3) professional, audit-ready expression (units, terminology, and thresholds that follow domain conventions).Detailed definitions and computation formulas of the evaluation metrics are included in Appendix A.5

In addition, as shown in figure 6, we conduct JudgeLM [27] blind pairwise comparisons: for each query, two systems' outputs are judged head-to-head, we swap left/right positions to reduce order bias, and average the two outcomes. We report three preference metrics: *WR* (ties excluded). Across most strong baselines, AgriGPT-VL achieves consistently higher win rates and remains competitive against top large models, corroborating the above advantages from a preference perspective. Appendix A.3 describes the prompt design for the LLM-based judger, and Appendix A.4 details the metric computation methodology.

Table 3. Ablation study of alignment stages. **Bold** indicates best per column.

Setting	Acc	Acc^+	BLEU	Meteor	Rouge-1-f	Rouge-2-f	Rouge-L-f
Base (Qwen2.5-VL-7B)	77.20%	60.32%	13.42	38.24	35.52	10.78	32.73
+ Shallow Alignment	78.23%	62.47%	15.54	36.47	40.76	14.07	37.95
+ Shallow + Deep	81.18%	66.67%	21.68	44.38	43.04	15.62	40.37
+ Shallow + Deep + GRPO	**85.84%**	**74.17%**	**26.27**	**47.55**	**46.52**	**20.09**	**43.81**

Table 4. Evaluation of general capabilities before and after fine-tuning.

Model	MMLU	ARC	OpenBookQA	MMBench	MMMU	SeedBench
Qwen2.5-VL-7B	0.6783	0.9043	0.8501	0.8398	0.4329	0.7565
AgriGPT-VL(7B)	0.6741	0.8462	0.8412	0.8312	0.4599	0.7574

4.2. Ablation Study

Starting from a base model, we progressively add *Shallow Alignment* (caption-only supervision with the vision stack frozen to establish cross-modal semantic anchors), *Deep Alignment* (single-choice reasoning with the vision encoder and cross-modal interaction layers unfrozen), and *GRPO* (reinforcement optimization with 15k GRPO reinforcement learning dataset).

As shown in tables 3, the results reveal a clear hierarchy of contributions: Shallow Alignment primarily improves lexical and descriptive consistency, stabilizing image–text keypoint correspondence; Deep Alignment is the main driver of cross-modal understanding and reasoning, lifting both discriminative and generation metrics; and GRPO further enhances factual faithfulness and robustness, with the largest gains on the stricter image-level cross-consistency metric (Acc^+), indicating that expert-level instructions are necessary to constrain high-precision behavior.

4.3. Generalization Evaluation

To assess whether domain specialization preserves general capabilities, we compare the fine-tuned *AgriGPT-VL* with its base model (*Qwen2.5-VL*) on six public benchmarks: three text-only (MMLU [25], ARC [13], OpenBookQA [35]) and three vision–language (MMBench [33], MMMU [60], SEED-Bench [29]).

Overall, *AgriGPT-VL* remains competitive. On text-only tasks, performance is largely preserved on MMLU and OpenBookQA, with only a modest decline on ARC. On vision–language tasks, the model matches or exceeds the base, showing parity on SEED-Bench and MMBench, and clear gains on MMMU.

As shown in Table 4, two conclusions emerge: (i) the curriculum—starting with textual grounding—effectively mitigates forgetting, maintaining broad competence across ∼40K out-of-domain samples; (ii) the improvement on MMMU confirms that learned visual reasoning generalizes beyond agriculture, reinforcing the strength and transferability of our finetuning framework.

4.4. External Evaluation on AgriBench

To further assess the robustness of our model, we conduct an additional evaluation on AgriBench[61], an external benchmark comprising approximately 700 single-choice questions. Unlike our in-house AgriBench-VL-4K, AgriBench differs entirely in data sources, annotation protocols, textual style, and question distribution, and is therefore fully decoupled from our training data and data-generation pipeline. This provides a genuinely out-of-distribution setting for examining whether the model can transfer across construction paradigms. Such an external evaluation not only eliminates potential biases arising from stylistic alignment but also offers a more comprehensive measure of the model's ability to interpret real agricultural scenarios.

As shown in tables 5, AgriGPT-VL achieves the highest accuracy of (70.10%) on AgriBench[61], clearly outperforming publicly available vision–language models including Gemini-2.5-Pro (56.94%). The significance of this independent evaluation lies in demonstrating that the improvements of AgriGPT-VL do not rely on the stylistic patterns of our training data nor on overfitting to a single construction pipeline. Instead, the model exhibits stronger grounding in agricultural knowledge, pest and disease characteristics, and field-level visual cues. For agriculture, achieving clear gains on a fully external benchmark indicates robust cross-source generalization.

5. Conclusion

We present AgriGPT-VL, an agricultural vision–language understanding suite that unifies large-scale data generation, curriculum-based multimodal training, and benchmark evaluation. The model demonstrates strong agronomic reasoning and visual grounding without sacrificing general ca-

Table 5. Model accuracy on AgriBench dataset.

Model	Acc
InternVL-3-8B	65.66%
InternVL-3-14B	60.08%
LLaVA-1.5-7B	59.29%
Qwen2.5-VL-7B	49.94%
Qwen2.5-VL-72B-Instruct	55.64%
Qwen/QVQ-72B-Preview	45.39%
DeepSeek-VL2	50.84%
Gemini-2.5-Flash	53.70%
Gemini-2.5-Pro	56.94%
AgriGPT-VL (ours)	**70.10%**

pabilities. This compact and reproducible framework provides a practical blueprint for building specialized multimodal systems in agriculture and beyond.

References

[1] 01.AI. Yi-vl-34b. Model card on Hugging Face, 2024. 6

[2] 01.AI. Yi-vl-6b. Model card on Hugging Face, 2024. 6

[3] Ali Abdallah. Plantvillage dataset. https://www.kaggle.com/datasets/abdallahalidev/plantvillage-dataset, 2019. Accessed: 2025-09-21. 3

[4] Josh Achiam and et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[5] Muhammad Arbab Arshad, Talukder Zaki Jubery, Tirtho Roy, Rim Nassiri, Asheesh K Singh, Arti Singh, Chinmay Hegde, Baskar Ganapathysubramanian, Aditya Balu, Adarsh Krishnamurthy, et al. Leveraging vision language models for specialized agricultural tasks. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6320–6329. IEEE, 2025. 3

[6] M. Awais et al. Agrogpt: An agricultural large language model. *arXiv preprint arXiv:2503.XXXX*, 2025. 2

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 6

[8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. Covers 3B/7B/72B; incl. 72B-Instruct. 6

[9] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. 6, 7

[10] Ling Chen and et al. Are we on the right way for evaluating large vision–language models? *arXiv preprint arXiv:2403.20330*, 2024. 2

[11] Xin Chen et al. Internvl: Scaling up vision-language learning and evaluation. *arXiv preprint arXiv:2404.14966*, 2024. 2

[12] Jennifer Clapp. *Food (3rd ed.)*. Polity, 2020. 2

[13] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. 8

[14] SelectDataset Contributors. Selectdataset: Crops and weeds. https://www.selectdataset.com/dataset/c2a3bab34be4ac974d93ffd1f7bbb39f, 2021. 558,930 images, 2,958 classes. Accessed: 2025-09-21. 3

[15] Densu341. Fresh and rotten fruits dataset. https://huggingface.co/datasets/Densu341/Fresh-rotten-fruit, 2022. 18 classes. Accessed: 2025-09-21. 3

[16] Raghav Didwania et al. Agrillm: Domain-specific large language models for agriculture. *arXiv preprint arXiv:2404.XXXX*, 2024. 2, 3

[17] Yilun Du et al. A survey of vision–language pre-trained models. *IJCAI Proceedings*, pages 5408–5415, 2022. 2

[18] Shenggen Fan and Christopher Rue. The role of smallholder farms in a changing world. In *The Role of Smallholder Farms in Food and Nutrition Security*, pages 13–28. Springer, 2020. 2

[19] Jonathan A. Foley, Navin Ramankutty, Kate A. Brauman, Emily S. Cassidy, James S. Gerber, Matt Johnston, and et al. Solutions for a cultivated planet. *Nature*, 478(7369):337–342, 2011. 2

[20] Aruna Gauba, Irene Pi, Yunze Man, Ziqi Pang, Vikram S Adve, and Yu-Xiong Wang. Agmmu: A comprehensive agricultural multimodal understanding and reasoning benchmark. *arXiv preprint arXiv:2504.10568*, 2025. 3

[21] H. Charles J. Godfray, John R. Beddington, Ian R. Crute, Lawrence Haddad, David Lawrence, James F. Muir, Jules Pretty, Sherman Robinson, Sandy M. Thomas, and Camilla Toulmin. Food security: The challenge of feeding 9 billion people. *Science*, 327(5967):812–818, 2010. 2

[22] Google DeepMind & Google AI. Gemini 2.5 flash. Gemini API official models page, 2025. Official model listing / description. 6

[23] Google DeepMind & Google AI. Gemini 2.5 pro. Gemini API official models page, 2025. Official model listing / description. 6

[24] Chao He, Lei Wu, Bo Yuan, Yaqian Li, Hao Yu, and Mingkui Tan. Species196: A one-million semi-supervised dataset for fine-grained species recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4572–4582, 2023. 3

[25] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Inter-*

national Conference on Learning Representations (ICLR), 2021. ICLR 2021. 8

[26] David P. Hughes and Marcel Salathé. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015. 3

[27] Fangkai Jiao, Bosheng Ding, Tianze Luo, and Zhanfeng Mo. Panda llm: Training data and evaluation for open-sourced chinese instruction-following large language models. *arXiv preprint arXiv:2305.03025*, 2023. 7

[28] Aristides Kamilaris and Francesc X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018. 2

[29] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308, 2024. 8

[30] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 6, 7

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. LLaVA-1.5 (incl. 7B) technical report. 6

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. LLaVA. 2

[33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, pages 216–233. Springer, 2025. 8

[34] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. DeepSeek-VL series (e.g., v1.2). 6

[35] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018. 8

[36] Horea Murean and Mihai Oltean. Fruits-360 dataset, 2018. 127,612 images, 206 classes. Accessed: 2025-09-21. 3

[37] Umair Nawaz, Muhammad Awais, Hanan Gani, Muzammal Naseer, Fahad Khan, Salman Khan, and Rao Muhammad Anwer. Agriclip: Adapting clip for agriculture and livestock via domain-specialized cross-model alignment. *arXiv preprint arXiv:2410.01407*, 2024. 3

[38] OpenBMB Team. Minicpm-v 2.6. Model card on Hugging Face, 2024. Model card describing v2.6 specifics. 6

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 6, 7

[40] Qwen Team. Qvq: To see the world with wisdom. Official Qwen blog, 2024. Multimodal reasoning model built on Qwen2-VL-72B. 6

[41] Saed Rezayi, Zhengliang Liu, Zihao Wu, Chandra Dhakal, Bao Ge, et al. Agribert: Knowledge-infused agricultural language models for matching food and nutrition. In *IJCAI 2022*, 2022. 2, 3

[42] Johan Rockström, Line Gordon, Carl Folke, Mats Lannerstad, and et al. Sustainable intensification of agriculture for human prosperity and global sustainability. *Ambio*, 46(S1): 4–17, 2017. 2

[43] J. Samuel et al. Agrollm: Large language models for agricultural assistance. *arXiv preprint arXiv:2501.XXXX*, 2025. 2, 3

[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image–text models. In *NeurIPS Datasets and Benchmarks*, 2022. 2

[45] Risa Shinoda, Nakamasa Inoue, Hirokatsu Kataoka, Masaki Onishi, and Yoshitaka Ushiku. Agrobench: Vision-language model benchmark in agriculture. *arXiv preprint arXiv:2507.20519*, 2025. 3

[46] M. S. Swaminathan. Food security and sustainable development. *Current Science or related venue*, 2001. Classic perspective on food security and sustainability. 2

[47] Qwen Team. Qwen-vl and qwen-vl-chat: Large-scale vision-language models. *arXiv preprint arXiv:2308.12966*, 2024. 2

[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. 3

[49] Yiqun Wang, Fahai Wang, Wenbai Chen, Bowen Lv, Mengchen Liu, Xiangyuan Kong, Chunjiang Zhao, and Zhaocen Pan. A large language model for multimodal identification of crop diseases and pests. *Scientific Reports*, 15 (1):21959, 2025. 3

[50] Sjaak Wolfert, Lan Ge, Cor Verdouw, and Marc-Jeroen Bogaardt. Big data in smart farming – a review. *Agricultural Systems*, 153:69–80, 2017. 2

[51] Jing Wu, Xinyu Li, Qianning Wang, Zelin Liu, Hanxiao Sun, Jianming Zheng, Guodong Long, Jing Jiang, and Yi Yang. The new agronomists: Language models are experts in crop management. *arXiv preprint arXiv:2403.19839*, 2024. 2

[52] Shang-Fu Wu et al. Ip102: A large-scale benchmark dataset for insect pest recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 3

[53] Lian Yan, Haotian Wang, Chen Tang, Haifeng Liu, Tianyang Sun, Liangliang Liu, Yi Guan, and Jingchi Jiang. Agrieval: A comprehensive chinese agricultural benchmark for large language models. *arXiv preprint arXiv:2507.21773*, 2025. 3

[54] Bo Yang and et al. Agrigpt: A large language model ecosystem for agriculture. *arXiv preprint arXiv:2508.08632*, 2025. 2, 3, 6, 7

[55] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 6

[56] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 2

[57] Gwang-Hyun Yu, Le Hoang Anh, Dang Thanh Vu, Jin Lee, Zahid Ur Rahman, Heon-Zoo Lee, Jung-An Jo, and Jin-Young Kim. Vl-paw: A vision–language dataset for pear, apple and weed. *Electronics*, 14(10):2087, 2025. 3

[58] Piaofang Yu and Bo Lin. A framework for agricultural intelligent analysis based on a visual language large model. *Applied Sciences*, 14(18):8350, 2024. 3

[59] S. et al. Yu. Insect foundation dataset. https://uark-cviu.github.io/projects/insect-foundation/, 2020. 317,128 images, 38,867 classes. Accessed: 2025-09-21. 3

[60] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024. 8

[61] Yutong Zhou and Masahiro Ryo. Agribench: A hierarchical agriculture benchmark for multimodal large language models. *arXiv preprint arXiv:2412.00465*, 2024. 8

[62] Hongyan Zhu, Shuai Qin, Min Su, Chengzhi Lin, Anjie Li, and Junfeng Gao. Harnessing large vision and language models in agriculture: A review. *arXiv preprint arXiv:2407.19679*, 2024. 3

[63] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. Covers InternVL-3 family, incl. 3-8B/3-14B. 6