Lychee-Bench: A Process-Level Vision Language Benchmark for Trustworthy Lychee Orchard Diagnosis

Guanyu Zhu[a,1], Kunming Lv[b,2], Hanbing Liu[a,c,3], Fengqi Yu[a,4], Lin Hu[a,5], Ersheng Ni[d,6], Meiyi Lu[a,7], Weixing Wang[e,f,8] and Jiaxing Xie[a,*]

[a] *College of Artificial Intelligence and Low-Altitude Technology, South China Agricultural University, Guangzhou 510642, China*
[b] *College of Engineering , South China Agricultural University, Guangzhou 510642, China*
[c] *National Center for International Collaboration Research on Precision Agricultural Aviation Pesticides Spraying Technology (NPAAC), South China Agricultural University, Guangzhou 510642, China*
[d] *The University of Queensland, Brisbane, Australia*
[e] *Zhujiang College, South China Agricultural University, Guangzhou 510900, China*
[f] *Guangdong Engineering Technology Research Center for Artificial Intelligence Empowered Media Technologies and Applications,Guangzhou 510900,China*

ARTICLE INFO

Keywords:
Precision agriculture
Lychee-Bench
Lychee-Mind
CoT-VQA

Abstract

Lychee is a high-value crop that is vulnerable to pests and diseases, where reliable diagnosis still depends on human experts. Current automated methods, such as CNN-based detectors, struggle to provide interpretability, while general-purpose vision-language models often hallucinate in lychee-growing agricultural environments. To address this, we collect high-resolution field images to construct the Lychee-dataset, featuring structured CoT-QA samples co-generated by large language models and verified by experts. We then develop Lychee-Mind, a domain-specific model trained via a two-stage paradigm that combines SFT with GRPO. By leveraging group-based comparisons to reward logical consistency without requiring a complex value model, Lychee-Mind internalizes expert reasoning patterns. We further evaluate our model on Lychee-Bench, a comprehensive benchmark comprising 12 fine-grained tasks ranging from insect identification to climate-aware causality analysis. Lychee-Mind achieves 80.67% accuracy, outperforming state-of-the-art open-source models by a significant margin. Ablation studies confirm that explicit CoT supervision and GRPO-based alignment are critical for process-aligned rewards. Overall, Lychee-Dataset, Lychee-Bench, and Lychee-Mind together provide a reproducible pipeline for more accurate and interpretable visionlanguage diagnosis in precision agriculture.

1. Introduction

Lychee (Litchi chinensis Sonn.) is a high-value fruit crop in tropical and subtropical regions, with an annual global output exceeding 3.5 million tonnes (Food and Agriculture Organization of the United Nations (FAO), 2023). Lychee has a long growing season and a complex management cycle: from flushing and flowering to fruit set and ripening, trees are vulnerable at multiple stages to a range of pests and diseases, including anthracnose, downy mildew, and fruit borers (Chen et al., 2024a). Its diagnosis still relies on experienced agronomists, who infer pest and disease status by visually inspecting leaves, fruits, and canopies and mentally integrating these cues with phenological stage and recent weather. This expert-dependent method is time-consuming and labour-intensive, and is becoming difficult to sustain under conditions of rural labour shortages, fragmented smallholder orchards, and more frequent extreme weather events.

The rapid growth of digital agriculture and intelligent sensing has prompted extensive research that employs computer vision techniques to support pest and disease monitoring in lychee and other fruit trees (Lan et al., 2025). Most existing studies formulate the problem as image classification and object detection, and improve performance through network architecture design, feature representation, or inference efficiency (Shengde et al., 2025; Li et al., 2022; Yin et al., 2025). Transformer- and CNN-based lesion recognition networks, lightweight embedded detection models,

xiaoguanbx@gmail.com (G. Zhu); lkmlkm7777777@gmail.com (K. Lv); nicelhb0118@163.com (H. Liu); 546766328@stu.scau.edu.cn (F. Yu); lynnhoo@stu.scau.edu.cn (L. Hu); ersheng.ni@student.uq.edu.au (E. Ni); lumeiyi640@stu.scau.edu.cn (M. Lu); weixing@scau.edu.cn (W. Wang); xjx1998@scau.edu.cn (J. Xie)

and lychee-specific detectors that use GAN-based data augmentation, modified FCOS backbones, and deformable-attention DETR variants to handle small targets and complex backgrounds (Wang and Xiao, 2021; Xie et al., 2023; Xiao et al., 2023; Wang et al., 2023, 2025). These methods achieve high accuracy in controlled settings and on single-disease benchmarks.Despite these advances, they have several limitations: First, most pipelines focus on static single-image inputs and single-task setups, making it difficult to cover the diagnostic process from early symptom onset to severe plant damage and to incorporate information about management history and growth stage (Zhou et al., 2025). Second, they expose only final labels and detection results (e.g., "healthy / diseased" labels, confidence scores, and bounding boxes), without explicit reasoning steps and evidence trail, practitioners cannot see why a given diagnosis is made or reuse it as a decision support (Li et al., 2025). Third, agricultural imagery is characterised by complex scenes, variable illumination, fragmented samples, and the need for expensive expert annotation, so high-quality, multi-task supervision signals remain scarce (Fang et al., 2024; Yang et al., 2025). Some recent efforts have begun to incorporate multimodal information in agricultural settings, for example, pairing images with textual descriptions and metadata. The associated annotations are still limited to coarse labels and short captions, and rarely provide explicit supervision signals over intermediate diagnostic reasoning. This makes it challenging to train models that can move beyond recognition to explainable, reasoning-driven decision support.

The rapid development of Large Multimodal Models (LMMs) has provided new insights for agricultural diagnosis. Early models mostly focused on open-domain vision-language tasks (e.g., LLaVA, BLIP, MiniGPT-4) , and subsequent researchers began to explore their adaptability in agricultural scenarios. Agri-LLaVA (Wang et al., 2024) enables the model to perform fine-grained pest identification and symptom explanation through the integration of agricultural pest knowledge graphs; AgMMU (Gauba et al., 2025) constructs a comprehensive benchmark covering disease diagnosis, yield estimation, and sustainability analysis, which is used to systematically evaluate the reasoning and generalization capabilities of LMMs; AgroGPT (Awais et al., 2024)improves the models performance in crop health assessment and agricultural consultation through domain expert tuning. These works demonstrate the potential of LMMs in domain knowledge injection, vision-language fusion, and open-ended question answering. However, existing agricultural multimodal benchmarks, such as Agri-Bench (Zhou et al., 2024)and AgMMU, place more emphasis on macro visual understanding across crops and multiple scenarios, while lacking fine-grained reasoning scenarios for specific cash crops and generally failing to include structured supervision of the reasoning process. For complex diagnostic tasks involving lesion evolution analysis, maturity inference, or symptom-yield correlation judgment, the absence of process-level annotations limits the credibility and interpretability of models in high-risk agricultural decision-making (Huang et al., 2025). In parallel with recognition-oriented visual methods, Large Multimodal Models (LMMs) have opened a new research avenue for agricultural diagnosis. General-purpose visionlanguage models (e.g., LLaVA, BLIP, MiniGPT-4) demonstrate strong capabilities in natural image understanding, question answering, and visual reasoning, and recent studies have begun to adapt such models to domains including remote sensing, medical imaging, and crop monitoring (Wang et al., 2024; Gauba et al., 2025; Awais et al., 2024). Methods such as Agri-LLaVA, AgMMU, and AgroGPT integrate domain knowledge graphs, curated benchmarks, and expert-tuned instructions to support tasks such as pest identification, disease diagnosis, yield estimation, and consultation. These efforts show that LMMs produce more flexible, interactive outputs than conventional detection models by injecting domain knowledge, fusing vision and language. However, these LMM-based approaches still primarily evaluate models based on final-answer correctness rather than whether their intermediate reasoning is consistent with visual evidence and agronomic knowledge. In fine-grained plant protection scenarios, general LMMs tend to provide plausible but incorrect descriptions and diagnoses when confronted with subtle lesion patterns or cluttered orchard backgrounds. Even when models provide correct answers, the lack of a reasoning process limits their usefulness for traceability and accountability in high-risk settings (Huang et al., 2025). Work on process supervision and reinforcement learning for language models has shown that aligning models with chain-of-thought (CoT) annotations and process-level rewards can improve both robustness and interpretability of generated reasoning, but there has been little exploration of such process-aligned training in multimodal agricultural diagnosis.

The above analysis reveals three key gaps for trustworthy diagnosis in lychee and other high-value crops. On the data side, there is still no large-scale, lychee-oriented visionlanguage dataset that covers multiple growth stages and viewpoints and provides questionanswer annotations that are aligned with agronomic concepts and step-by-step diagnostic reasoning. On the evaluation side, current agricultural benchmarks lack a systematic, multidimensional testbed that jointly probes perception, analysis, diagnosis, decision-making, and reflective error-checking in realistic orchard scenarios. On the modelling side, there is a lack of process-aligned training frameworks tailored to high-risk agricultural diagnosis: recognition-oriented visual models expose only final predictions rather than their reasoning traces, while existing agricultural LMMs are mainly trained and evaluated on final-answer correctness.

To address these challenges, we propose a holistic framework for trustworthy lychee orchard diagnosis. First, we construct Lychee-Dataset, a high-quality dataset of structured “Question-Thinking-Answer” samples generated by large language models(LLMS) and subsequently screened and refined by agronomy experts, providing the process-level supervision that is missing in existing agricultural multimodal datasets. Second, we establish Lychee-Bench, a hierarchical benchmark that decomposes diagnosis into 12 fine-grained tasks, ranging from symptom localization to climate-aware causality analysis. Finally, built upon the Qwen2.5-VL-7B foundation, we introduce Lychee-Mind, a domain-specific LMM trained through a two-stage framework: Supervised Behavior Initialization + Group Relative Policy Optimization(GRPO) Process Alignment. In the first stage, we apply Supervised Fine-Tuning (SFT) on structured CoT trajectories from Lychee-Dataset to embed the expert's "Question-Thinking-Answer" cognitive pattern, creating a core behavioral prior for diagnostic reasoning. In the second stage, we implement GRPO to explicitly assess the quality of the evidence and the consistency between the evidence and the conclusions.

Overall, this work makes the following contributions:

1) We construct the **Lychee-Dataset** dataset, a large-scale lychee visionlanguage corpus in which each sample is annotated with a three-stage QuestionThinkingAnswer structure, providing process-level supervision that compensates for the lack of intermediate reasoning annotations in existing agricultural multimodal datasets.
2) We propose a two-stage training framework for the **Lychee-Mind** model that combines supervised fine-tuning for behaviour initialization with GRPO-based reinforcement learning. We design a process-level reward based on evidence aggregation to optimise the entire reasoninganswer trajectory, enabling the model to generate thinking chains that are consistent with image evidence and agronomic knowledge while maintaining accurate diagnostic conclusions.
3) We introduce **Lychee-Bench**, a comprehensive evaluation benchmark that organises Lychee-Dataset into multi-level agricultural visual reasoning tasks, covering cognitive dimensions such as perception, understanding, analysis, diagnosis, decision making, reflection, and generation. Lychee-Bench provides a standardised and reproducible framework for evaluating reasoning capabilities in Lychee-related scenarios.

2. Materials and Methods

2.1. Data acquisition

To support the proposed two-stage training framework, we constructed a large-scale multimodal dataset for lychee disease diagnosis, pest infestation assessment, and maturity evaluation, referred to as the Lychee-Dataset. All data were collected in real lychee production orchards located in Guangzhou, China (23°16′N, 113°36′E), using ground-based close-range field imaging. Handheld digital cameras and high-resolution DSLR devices were employed under a multi-scale and multi-view acquisition strategy, enabling the capture of hierarchical visual information ranging from localized symptoms to global canopy structure. To provide an intuitive overview of the geographical sources and visual characteristics of the dataset, Figure 1 visualizes the orchard distribution, data collection sites, and representative field images that illustrate typical disease, pest, and maturity conditions contained in the Lychee-Dataset.

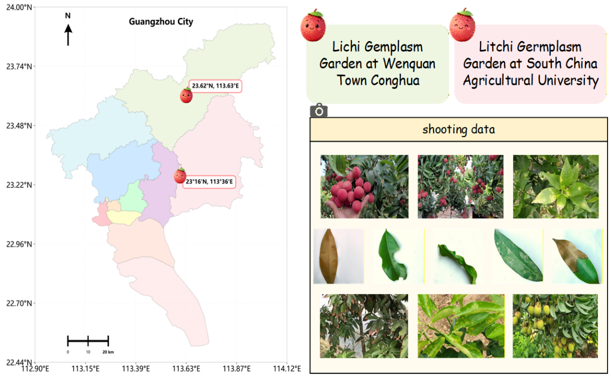

Figure 1: Overview of field data acquisition. (a) Geographical distribution of the two lychee orchards in Guangzhou, China, where all field images were collected. (b) Illustration of the two data collection sites: the Lychee Germplasm Garden at Wenquan Town, Conghua, and the Lychee Germplasm Garden at South China Agricultural University. (c) Representative multi-scale shooting samples, including close-up fruit images, leaf-level disease and pest symptoms, and whole-tree or canopy-level field scenes.

This data acquisition strategy spans the full temporal progression of lychee fruit developmentfrom the early enlargement stage, through peel coloration and flavor formation, to the peak harvest period and the late-season phaseall-owing the dataset to capture continuous visual transitions in morphology, surface color, and fruit texture. In terms of phenotypic diversity, the dataset encompasses multi-scale field scenarios ranging from individual fruits and small fruit clusters to whole-tree canopies and localized leaf or branch regions. It includes both healthy specimens with normal growth status and a wide spectrum of pathological conditions, covering latent stress or early asymptomatic infection, subtle discoloration and initial lesion formation, intermediate stages with evident disease symptoms, rot, or pest chewing marks, and severe late-stage manifestations such as extensive necrosis, wilting, and heavy pest damage on fruits, leaves, and branches.

All imaging devices were configured to enable or reserve RAW-format output to preserve the original scene's dynamic range and color fidelity, while simultaneously generating JPEG images to support diverse downstream tasks. Figure 2(a) presents the hierarchical composition of the dataset using a concentric sunburst structure, where categories expand progressively from primary classes to fine-grained phenotypes, reflecting the natural organization of lychee diseases, pests, and maturity traits and enabling flexible task partitioning based on semantic hierarchy. Figure 2(b) illustrates the distribution of images and QA samples across the three major categoriesdisease incidence, pest infestation, and fruit maturitywith disease-related samples accounting for the largest proportion.

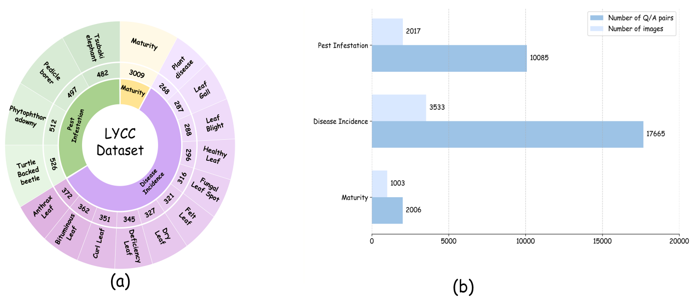

Figure 2: Overview of the Lychee-Dataset. (a) Hierarchical composition of the dataset. The sunburst chart shows the fine-grained category distribution, where the inner ring represents the three primary categoriesDisease Incidence, Pest Infestation, and Maturityand the outer ring displays their corresponding sub-classes. (b) Quantitative statistics of the dataset. The bar chart reports the number of images and questionanswer pairs within each primary category.

Overall, the data acquisition process was designed to fully account for real production conditions, including variations in imaging environments, observation scales, and temporal development. This ensures that the resulting dataset provides a structurally complete, visually realistic, and phenotype-rich foundation for downstream tasks such as lychee maturity assessment and diseasepest diagnosis.

2.2. Data processing

To address the lack of process-level supervision in agricultural diagnosis, we adopt a process-oriented design for the Lychee-Dataset that explicitly encodes step-by-step diagnostic reasoning.Based on 6,553 high-resolution field images, we construct in total 29,756 structured Chain-of-Thought Question Answering (CoT-QA) samples, replacing traditional static "question-answer" annotations with fine-grained modelling of the reasoning process. Figure 3 illustrates the overall pipeline for constructing CoT-QA samples, including image preprocessing, question generation, COT generation, and expert review.

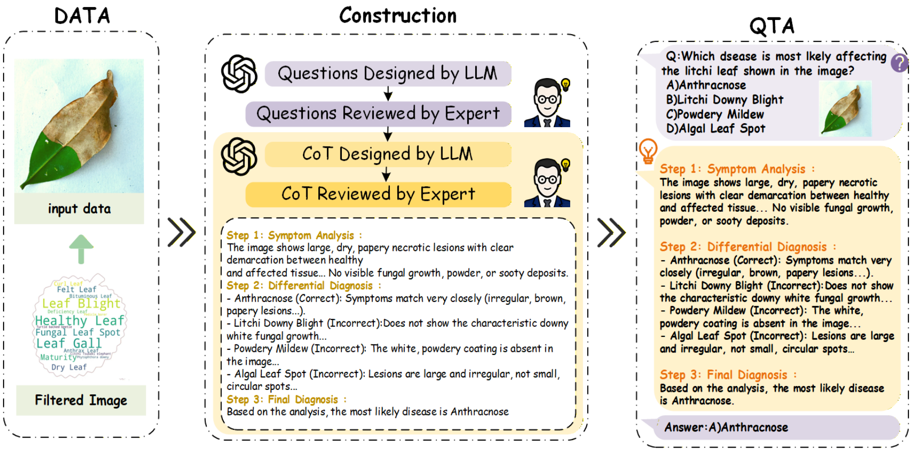

Figure 3: Overview of the CoT-QA construction pipeline in the Lychee-Dataset. The workflow includes image preprocessing, expert-guided question design, large-language-model-generated reasoning chains, and expert verification, resulting in structured Question–Thinking–Answer samples for process-level supervision.

Specifically, we adopt a hybrid pipeline that combines expert-designed templates, large-language-model-based generation, and expert review. Agricultural experts first analysed typical scenarios in lychee maturity assessment and diseasepest diagnosis, and then defined standardised question templates and key diagnostic criteria. Under these template constraints, a large language model was employed to generate large-scale QA texts. The generated questions, reasoning chains, and answers were then checked and refined by domain experts to remove logically inconsistent or agronomically implausible cases and to unify the diagnostic criteria. Each sample was ultimately organised into a structured record containing the image ID, scene type, fine-grained category, task type, standardised question and options, the corresponding chain-of-thought reasoning process, and a single correct answer. Additional fields such as difficulty level and disease progression stage were included to support programmable dataset partitioning by scene, task type, or complexity.

Beyond conventional plant pathology datasets that only provide imagelabel pairs, the Lychee-Dataset explicitly encodes process-level supervision in the Thinking segment. All samples are standardised into a three-stage QuestionThinkingAnswer format, where the core supervision signal is the chain-of-thought (CoT) stored in the Thinking field.

The Thinking stage requires the model to conduct a structured analysis of each case before committing to a diagnosis. It typically involves: (i) describing local visual evidence such as lesion location, colour and morphology, pest body traits, and the condition of fruits and leaves; (ii) incorporating background information from the question stem (e.g., cultivar, season, and cultivation practices) into causal reasoning; and (iii) comparing and discriminating between candidate diagnoses. Together, these steps define a reasoning trajectory that links symptom observation, causal analysis, option comparison, and final decision. Under real field conditions, the process-level annotations therefore provide not only the correct label but also the expert step-by-step diagnostic reasoning, which later serves as the supervision signal for process-aligned training.

Building on this corpus of QuestionThinkingAnswer (QTA) trajectories, we use it as the unified supervision source for the two-stage training of Lychee-Mind. In essence, each QTA sample can support the two stages in complementary ways. For Stage 1, we allocate the full set of 29,756 QTA samples as training targets, so that the model learns, given the image and question, to produce a coherent ThinkingAnswer reasoning process and the final decision. For Stage 2, we further construct an RL pool of 5,000 cases sampled from the same training split: for each case, the imagequestion pair is kept fixed as context, and the policy generates multiple alternative ThinkingAnswer continuations that are scored by a process-level reward model under the GRPO framework. Figure 4 summarises how the Lychee-Dataset is converted into training sets for the Stage 1 SFT and Stage 2 RL stages, while the corresponding learning objectives are detailed in Sections 2.3 and 2.4.

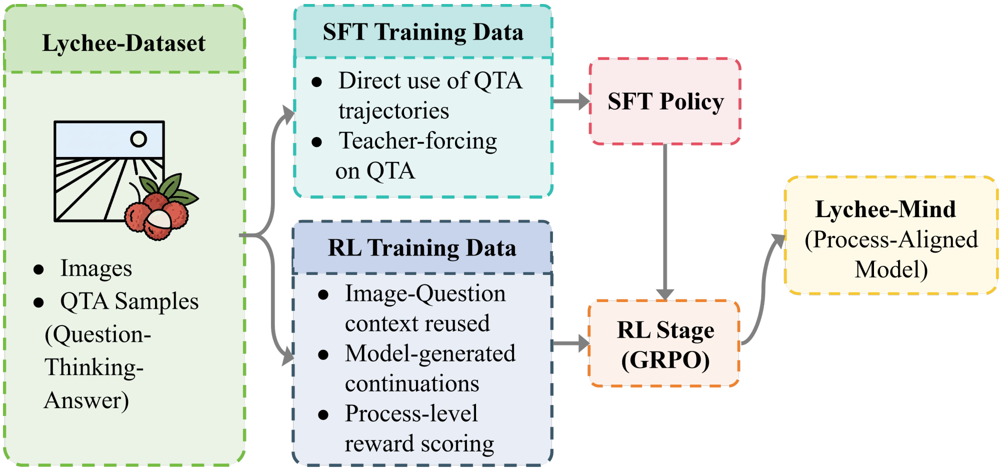

Figure 4: Use of Lychee-Dataset in the two-stage training pipeline. The same Question–Thinking–Answer (QTA) samples are used directly for supervised fine-tuning (top path) and, after generating multiple candidate reasoning–answer trajectories and scoring them with a reward model, for GRPO-based reinforcement learning (bottom path), yielding the final process-aligned Lychee-Mind policy.

2.3. Behavioral Initialization

2.3.1. Formalization of supervised trajectories

Building on the QTA trajectories defined in Section 2.2, we now view each annotated case as a supervised inputoutput pair for sequence modelling. Each sample provides an imageinstruction input together with a complete ThinkingAnswer trajectory, and our goal in this stage is to learn a policy that maps the former to the latter in an autoregressive manner. We introduce the following notation.

Let the input be

$$X = (I_{\text{img}}, I_{\text{ins}})$$

where I_{img} denotes the lychee-related image, and I_{ins} denotes the textual instruction that includes the question and candidate options. The corresponding output is a textual trajectory

$$Y = (y_1, y_2, \dots, y_L)$$

This trajectory contains the complete ThinkingAnswer segment, i.e., all tokens from the reasoning process to the final answer. To explicitly distinguish between the reasoning process and the decision outcome, we further decompose the output as

$$Y = [Y_{\text{think}}; Y_{\text{ans}}]$$

where Y_{think} denotes the token sequence corresponding to the chain-of-thought (CoT) evidence, and Y_{ans} denotes the token sequence corresponding to the final single-choice answer. The construction procedure and specific content of the chain-of-thought have been detailed in Section 2.2; here we treat it as the expert-level diagnostic or assessment process that the model is expected to approximate during training.

Let the supervised dataset be

$$D_{\text{sft}} = \{(X_i, Y_i)\}_{i=1}^{N}$$

and the subsequent supervised fine-tuning stage uses this dataset to initialize the model's behavior.

2.3.2. Supervised fine-tuning for behavioral initialization

Given the supervised dataset D_{sft}, we employ standard supervised fine-tuning (SFT) to initialize the policy model. Let θ denote the set of trainable model parameters, including all parameters in the vision encoder, the language model, and the multimodal projection layers. The objective of the SFT stage is to maximize the autoregressive likelihood of the target trajectories, which is equivalent to minimizing the empirical negative log-likelihood with a regularization term. The overall optimization objective is

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \left(\mathcal{L}_{\text{SFT}}(\theta) + \lambda R(\theta) \right),$$

where θ^* denotes the model parameters obtained after SFT, $\mathcal{L}_{\text{SFT}}(\theta)$ is the main loss defined on task-related data, λ is the regularization coefficient, and $R(\theta)$ is a regularization term (e.g., the L2 norm of the parameters) used to mitigate overfitting and improve generalization across diverse field conditions.

To illustrate how the model is behaviorally initialized through chain-of-thought demonstrations in the first stage, we visualize the core workflow of the SFT process, as shown in Figure 5.

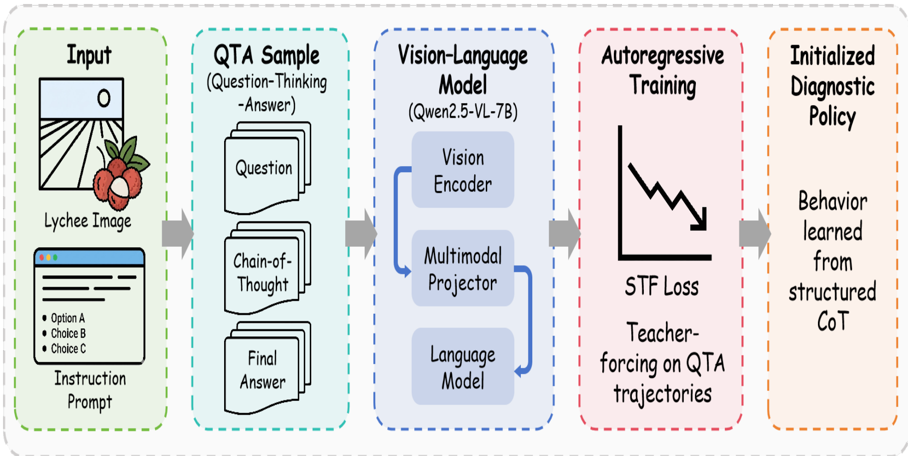

Figure 5: Illustration of the Stage-1 Supervised Fine-Tuning (SFT) process.

The main loss $\mathcal{L}_{\text{SFT}}(\theta)$ for a single supervised trajectory is defined as follows. Let $(X, Y) \in D_{\text{sft}}$ be a supervised sample, where X denotes the combined input comprising the lychee image and textual instruction (question and candidate options), and $Y = (y_1, \dots, y_L)$ is the corresponding target output trajectory that sequentially includes all tokens in the ThinkingAnswer segment. Let P_θ denote the conditional autoregressive distribution parameterized by θ. Then

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(X,Y)\sim D_{\text{sft}}} \left[\frac{1}{L} \sum_{t=1}^{L} \log P_\theta \left(y_t \mid X, y_{<t} \right) \right],$$

where $\mathbb{E}[\cdot]$ denotes the expectation under the supervised data distribution; L is the length of the target trajectory; y_t is the target token at position t; $y_{<t}$ denotes the prefix consisting of all tokens generated before position t; and $P_\theta \left(y_t \mid X, y_{<t} \right)$ denotes the conditional probability of generating token y_t given the input X and the historical prefix $y_{<t}$.

2.4. Reinforcement learning

2.4.1. Group Relative Policy Optimization for diagnostic trajectories

Supervised fine-tuning on Question–Thinking–Answer trajectories equips Lychee-Mind with the ability to imitate expert reasoning patterns on the training distribution. However, as discussed in the Introduction, such imitation alone is insufficient for trustworthy agricultural diagnosis: models may produce fluent but brittle explanations and may fail to generalise when encountering unseen symptom combinations, unfamiliar lighting conditions, or rare pest–disease co-occurrence cases. To address this limitation, the second stage introduces reinforcement learning on the same QTA corpus, as illustrated in Fig. 6, in order to improve both the robustness of the reasoning process and the model's generalisation beyond the supervised data distribution.

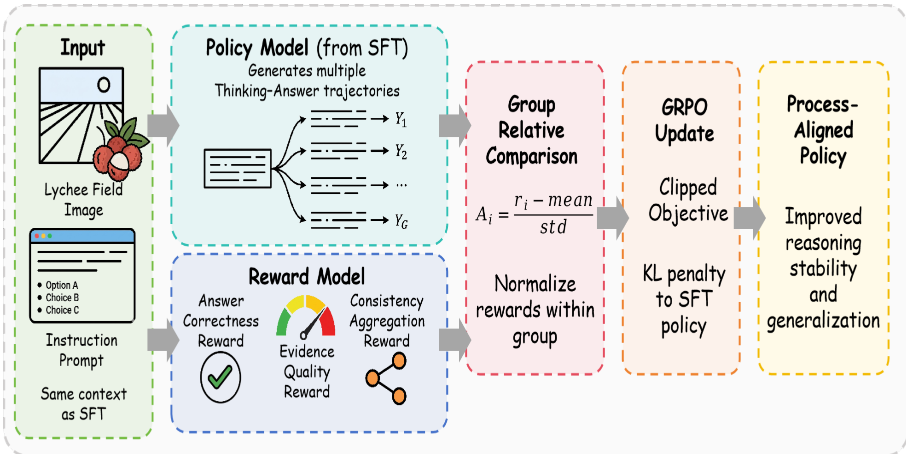

Figure 6: Overview of the GRPO-based reinforcement learning stage. For each lychee field image and instruction, the SFT-initialised policy samples multiple Thinking–Answer trajectories, which are scored by a process-level reward model; group-relative normalisation yields advantages that drive GRPO updates towards a process-aligned diagnostic policy.

We adopt Group Relative Policy Optimization (GRPO), a variant of Proximal Policy Optimization (PPO) that replaces an explicit value function with group-wise relative scoring. In our setting, the input $X = (I_{\text{img}}, I_{\text{ins}})$ already contains the question, and the learnable policy governs the continuation from "Thinking" to "Answer". Each sampled "Thinking–Answer" sequence is treated as a *diagnostic trajectory*, a shortened notation of the Question–Thinking–Answer structure defined in Section 2.2 where the question is fixed as context rather than a learnable segment.

Compared with standard PPO, GRPO is particularly suitable for our multi-dimensional process rewards and for enhancing generalisation ability. By comparing several trajectories generated from the *same* input case (e.g., multiple alternative diagnostic explanations for one field image), GRPO optimises the policy based on relative quality, rather than absolute scalar values that can be noisy or subjective. This ranking-based formulation mirrors how human experts judge diagnostic soundness across possible explanations, strengthens the influence of process-level rewards, and empirically leads to more robust behaviour on unseen orchard conditions.

Concretely, for each RL training sample X, the old policy $\pi_{\theta_{\text{old}}}$ samples G diagnostic trajectories from the same input:

$$\{Y_1, Y_2, \dots, Y_G\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid X),$$

and a scalar reward r_i is computed for each trajectory (see Section 2.4.2). Let the reward set for this input be

$$R = \{r_1, \dots, r_G\},$$

then GRPO constructs the advantage of each trajectory via group-wise standardization:

$$A_i = \frac{r_i - \text{mean}(R)}{\text{std}(R) + \epsilon},$$

where ϵ is a small positive constant (e.g., 10^{-8}) that prevents numerical instability when all trajectories receive very similar rewards. Intuitively, a trajectory is rewarded if it scores higher than its peers for the *same* case, even when the absolute reward is noisy.

Let the length of trajectory Y_i be L_i, and denote the token at position t by $y_{i,t}$. The probability ratio between the new and old policies at this token is

$$\rho_{i,t} = \frac{\pi_\theta\left(y_{i,t} \mid X, y_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(y_{i,t} \mid X, y_{i,<t}\right)}.$$

The GRPO objective follows a PPO-style clipped surrogate loss with an additional KullbackLeibler (KL) regularization term:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{L_i}\sum_{t=1}^{L_i}\left(\min\left(\rho_{i,t}A_i,\ \text{clip}\left(\rho_{i,t}, 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}}\right)A_i\right) - \beta\, D_{\text{KL}}\left(\pi_\theta \parallel \pi_{\text{ref}}\right)\right)\right],$$

where ϵ_{clip} is the clipping coefficient that controls the magnitude of policy updates, β is the KL regularization coefficient, and the reference policy π_{ref} is set to the SFT policy π_{sft} obtained in the first stage. In our setting, rewards are computed after the entire ThinkingAnswer trajectory has been generated, so all tokens within the same trajectory share the same advantage A_i.

This design removes the need for an additional value network, which is difficult to train reliably on subjective, process-level rewards, and instead focuses learning on relative improvements between competing diagnostic trajectories for the same orchard case.

2.4.2. Process reward function based on evidence aggregation

Within the GRPO framework described above, the reward attached to each diagnostic trajectory determines *which* behaviours are considered better. If we rely solely on an outcome reward based on answer correctness, e.g.,

$$r_{\text{answer}} = \mathbf{1}(Y_{\text{ans}} = \text{Ground Truth})$$

we fall back to the limitation discussed in the Introduction: the model may obtain high returns by guessing the correct answer while writing a coarse, shortcut-style chain-of-thought that does not carefully describe lesion location, pest traits, or growth stage. Such trajectories are unsafe for agricultural decision support because the final conclusion is not backed by explicit, verifiable evidence.

To address this problem, we introduce a process-level reward function that measures both the intrinsic quality of the evidence chain and its consistency with the final conclusion. The total reward

$$r_{\text{total}} \quad (\text{i.e., } r_i \text{ in the GRPO framework})$$

is defined as a weighted combination of three components: an evidence quality reward r_{evidence}, an answer correctness reward r_{answer}, and a consistency aggregation reward $r_{\text{aggregation}}$.

Evidence quality reward $r_{evidence}$. This term evaluates the agronomic soundness and logical structure of the Thinking part, regardless of whether the final option is correct. We employ an external large language model as a reward evaluator and specialize it to the agricultural diagnostic setting through carefully designed prompts and few-shot examples. Each prompt provides (i) the field image description and question, (ii) the candidate options, and (iii) a single chain-of-thought, and asks the evaluator to score the reasoning on criteria such as observational accuracy (e.g., whether the described symptoms actually match the image), causal plausibility, and relevance to the question. The score is then mapped to a scalar in $[-1, 1]$:

$$r_{\text{evidence}} = \text{Judge}\left(Y_{\text{think}}, X\right) \in [-1, 1].$$

This encourages the model to produce detailed, agronomically meaningful explanations rather than generic or template-like CoT.

Answer correctness reward r_{answer}. This term serves as the primary task signal and directly evaluates whether the final answer Y_{ans} matches the expert ground truth:

$$r_{\text{answer}} = \begin{cases} 1, & \text{if } Y_{\text{ans}} = \text{Ground Truth}, \\ 0, & \text{otherwise.} \end{cases}$$

It ensures that the model does not sacrifice diagnostic accuracy when optimizing process-level behaviour.

Consistency aggregation reward $r_{aggregation}$. The previous two terms alone do not yet enforce that good-looking reasoning must lead to the correct conclusion. In particular, increasing λ_2 in front of r_{evidence} would globally reward fluent CoT even when it supports a wrong option. To explicitly reward *evidenceconclusion consistency*, we therefore introduce

$$r_{\text{aggregation}} = \mathbb{I}\left(Y_{\text{ans}} = \text{Ground Truth}\right) \cdot r_{\text{evidence}},$$

where $\mathbb{I}(\cdot)$ is the indicator function. Only trajectories that are both correct *and* well reasoned receive additional positive reward. Intuitively, r_{evidence} encourages writing a high-quality diagnostic story, while $r_{\text{aggregation}}$ amplifies such stories only when they justify the right diagnosis, preventing the model from learning persuasive but wrong explanations.

Weighting and hyperparameter search. Finally, the total reward applied to the entire trajectory (i.e., the reward fed into GRPO as r_i) is defined as

$$r_i = r_{\text{total}} = \lambda_1 r_{\text{answer}} + \lambda_2 r_{\text{evidence}} + \lambda_3 r_{\text{aggregation}},$$

where $\lambda_1, \lambda_2, \lambda_3 \geq 0$ are weighting coefficients that balance the three terms. We treat answer correctness as the baseline scale and fix $\lambda_1 = 1.0$. The coefficients λ_2 and λ_3 are selected by grid search on the validation split of Lychee-Bench: we restrict the search to a small set of combinations with $\lambda_2, \lambda_3 \in [0, 1]$ and choose the pair that maximizes overall validation accuracy while keeping the average evidence score high. This procedure avoids ad-hoc manual tuning and makes the trade-off between answer accuracy and reasoning quality explicit.

By combining this structured reward function with the GRPO framework, the reinforcement learning stage constrains both the model's final predictions and its reasoning process. The model is encouraged not only to produce conclusions that match expert diagnoses, but also to generate evidence chains that are agronomically plausible, logically coherent, and tightly coupled to the chosen option. Compared with reward designs based solely on answer correctness, this process-aligned training paradigm yields more interpretable diagnostic trajectories and empirically improves robustness on complex or atypical disease cases, as shown in Section 3.

3. Experiments

3.1. Benchmark: Lychee-Bench

Lychee-Bench is a dedicated evaluation suite built on real orchard images and designed to measure both visual understanding and cross-task reasoning in lychee production scenarios. It organises end-to-end cognitive tasks into three major domainsPlant Disease, Pest Infestation, and Maturityand further decomposes them into twelve fine-grained task types. These include Pest/Disease Diagnosis (PD), Visual Prompt Reasoning (VPR), Anomaly Reasoning (AR), Climate & Environmental Reasoning (CTR), Agronomic Planning & Decision-making (APD), Pest Identification (PI), pest-oriented variants of VPR/AR/CTR/APD, maturity-related APD, and Maturity Stage Classification (MSC).

Together, these tasks cover a wide spectrum of cognitive demands, from basic symptom recognition and semantic understanding to multi-factor causal reasoning and scenario-based decision-making. Figure 7 summarises the organisation of Lychee-Bench by domains and task types, along with representative real-world examples.

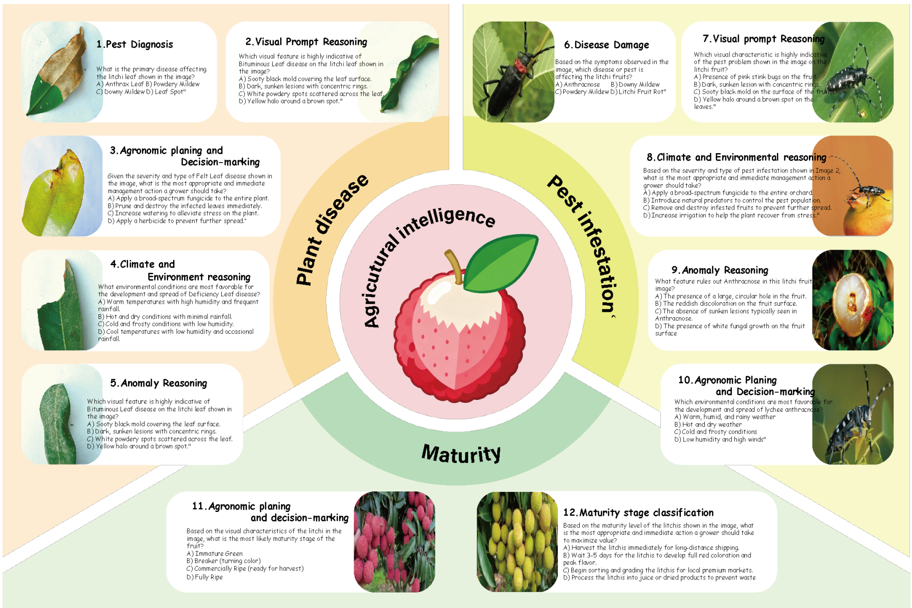

Figure 7: Taxonomy of the Lychee-Bench. The benchmark evaluates agricultural intelligence across three core scenarios: Plant Disease, Pest Infestation, and Maturity. Each scenario includes multiple fine-grained reasoning tasks, such as diagnosis, visual reasoning, and decision-making, presented in a multiple-choice format with illustrative examples.

3.2. Models and Baselines

To provide a comprehensive comparison across the 12 task types, we evaluate Lychee-Mind against 24 representative large multimodal models (LMMs), including 4 closed-source and 20 open-source systems. The closed-source models are GPT-4o, Gemini 1.5 Flash, Gemini 1.5 Pro, and Yi-VL-6B. The open-source baselines cover several mainstream model families: instruction-tuned models such as Qwen2-VL-7B and LLaVA-Next, mid-scale models including InternVL2, larger-capacity systems such as DeepSeek-VL2 and LLaVA-NeXT-Vicuna-13B, and lightweight models including Phi-3.5-Vision and MiniCPM-Llama3-V-2.5.

All open-source models are evaluated using their official checkpoints under a unified inference protocol, while all closed-source evaluations are conducted on a compute node equipped with four NVIDIA A100 GPUs.

3.3. Task performance comparison

This section provides a systematic analysis of the overall performance of Lychee-Mind on Lychee-Bench and conducts a comprehensive comparison with leading open-source and closed-source multimodal models. The discussion focuses on two key dimensions: (1) the relationship between model performance and parameter scale, and (2) performance across the 12 fine-grained task categories. Overall, Lychee-Mind demonstrates substantial advantages across multiple tasks, particularly in diagnostic scenarios that require deep domain understanding and robust reasoning capabilities.

Table 1 presents the detailed performance of all models across the sub-tasks. Although the accuracy of vision-language models typically increases with parameter size, this trend does not strictly hold in domain-specific agricultural tasks. Notably, the 7B-parameter Lychee-Mind not only surpasses models of similar scale but also outperforms several larger models, such as LLaVA-NeXT-Vicuna-13B. This demonstrates that the combination of domain-specific data construction, process-level supervision, and a two-stage training paradigm enables a highly efficient performance-to-parameter ratio even at moderate model scales. Compared with the base model Qwen2.5-VL-7B (67.50%), Lychee-Mind achieves a 13.17% improvement in overall accuracy, further confirming the effectiveness of our training framework in aligning domain knowledge and modeling reasoning chains. From the perspective of task-level performance, the advantages of Lychee-Mind are particularly prominent in core diagnostic tasks that require strong domain expertise.

Table 1
Detailed performance comparison of all models across the 12 sub-tasks in Lychee-Bench

Model	Plant Disease					Pest Infestation					Maturity		Overall
	PD	VPR	AR	CTR	APD	PI	VPR	AR	CTR	APD	MSC	APD	
Closed-source Models													
GPT-4o (OpenAI, 2024)	61	87	76	89	**93**	52	76	74	91	**87**	76	**94**	79.67
Gemini-1.5-pro (G.D. Team Gemini, 2024)	56	90	73	**96**	81	71	**91**	81	94	82	76	78	**80.75**
Gemini-1.5-flash (Team Gemini, 2024)	67	**91**	**80**	91	76	68	80	74	92	77	54	73	76.92
Open-source Models													
LLaVA-1.5-7B (Liu et al., 2024a)	57	71	71	92	54	63	88	**84**	94	56	54	71	71.75
LLaVA-NeXT-Video-7B (Liu et al., 2024b)	31	63	61	81	43	42	82	67	88	48	59	54	59.92
InternVL2-8B (Chen et al., 2024b)	41	80	61	89	64	45	87	65	92	79	75	71	70.75
Qwen2.5-VL-7B (Bai et al., 2025)	42	76	53	94	83	46	74	56	89	66	64	67	67.50
Yi-VL-6B (01-AI, 2024)	41	67	42	87	74	36	81	55	**97**	62	59	49	62.50
MiniCPM-Llama3-V 2.5 (Yao et al., 2024)	30	74	61	88	82	27	85	50	79	68	71	64	64.92
MiMo-VL-7B-RL-2508	**81**	76	76	92	36	**93**	83	78	83	33	58	41	69.17
LLaVA-Phi-3-mini	37	63	43	94	75	36	72	48	87	62	60	79	63.00
Qwen2-VL-2B-Instruct	32	67	62	87	63	29	57	68	81	32	63	62	58.58
Phi-3.5-Vision (Abdin et al., 2024)	38	79	73	92	74	44	87	61	83	65	67	71	69.50
LLaVA-NeXT-Vicuna-13B (Liu et al., 2024c)	71	81	76	91	81	72	67	74	95	79	74	73	77.83
Lychee-Mind	**74**	**85**	**79**	**95**	**82**	**86**	**81**	**74**	**92**	**68**	**78**	**74**	**80.67**

In the pest identification (PI) task, the model achieves 86% accuracy, substantially outperforming GPT-4o (52%) and the base model Qwen2.5-VL-7B (46%), demonstrating its ability to effectively integrate fine-grained visual cues with agricultural knowledge for systematic decision-making. Similarly, in the plant disease diagnosis (PD) task, Lychee-Mind attains 74% accuracy, exceeding most comparison models and indicating stronger robustness in recognizing complex disease symptoms.

In the more challenging cross-modal reasoning tasks, Lychee-Mind likewise demonstrates strong performance. For example, in the climate and environmental reasoning (CTR) task, the model achieves a high score of 95%, indicating its strong capability in handling multivariate causal relationships and integrating information across modalities. In the visual prompt reasoning (VPR) and anomaly reasoning (AR) tasks, the model's stable outputs further illustrate its enhanced consistency in reasoning and its improved ability to detect abnormal patterns.

Moreover, in the maturity stage classification (MSC) task, which requires fine-grained visual discrimination, Lychee-Mind leads with 78% accuracy, reflecting its sensitivity to subtle visual cues. Taken together, these results show that Lychee-Mind's performance gains stem not only from effective absorption of diagnostic knowledge but also from its holistic improvements in reasoning consistency, visual understanding, and cross-modal information integration.

Lychee-Mind exhibits a comprehensive and well-balanced capability across the Lychee-Bench, and its leading performance on multiple key tasks demonstrates that the combination of process-level supervision, reward-based optimization, and domain-specific data enables deep understanding and reliable reasoning for specialized agricultural tasks, even under a relatively small parameter scale.

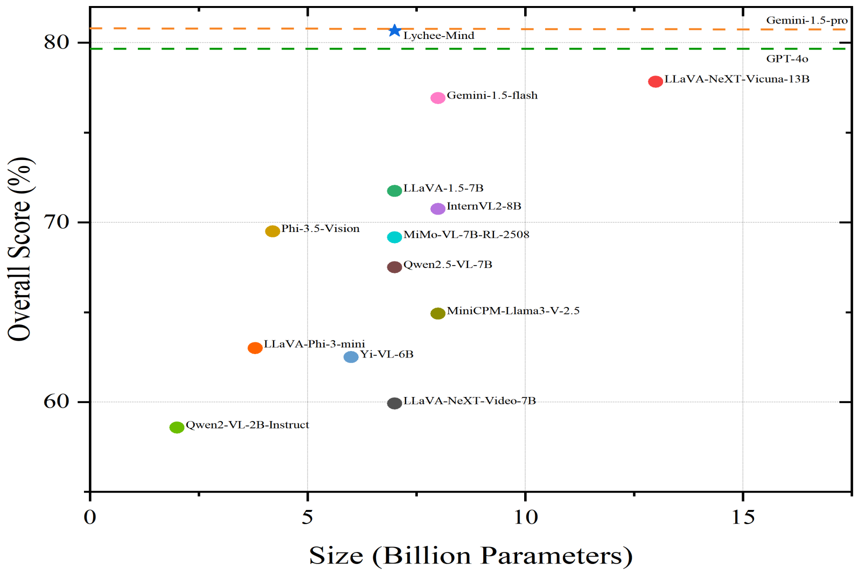

Figure 8: Performance comparison across parameter scales on the Lychee-Bench .

3.4. Ablation studies

To systematically assess the independent contribution of each key component in the Lychee-Mind training framework and to gain deeper insight into the models reasoning mechanisms in professional agricultural tasks, a series of ablation experiments were conducted. These experiments are organized along two complementary dimensions. First, we compare three training strategiessupervised fine-tuning (SFT), SFT + PPO, and SFT + GRPOto evaluate their respective impacts on visual understanding and cross-task diagnostic performance, thereby validating the role and performance gains of each strategy within the proposed two-stage training paradigm. Second, we study the effect of data structure by contrasting Flat QA with QA + CoT, examining how different annotation formats influence reasoning-chain quality, causal consistency, and final diagnostic outcomes. Together, these experiments reveal fine-grained behavioral differences in how the model performs under complex and high-risk agricultural scenarios and shed light on the stability and reliability of its reasoning generation process.

3.4.1. Ablation on training strategies

To systematically evaluate the impact of different training strategies on agricultural visual reasoning tasks, we constructed three configurations based on a unified model architecture: (1) SFT only: supervised fine-tuning on the full-chain Lychee-Dataset; (2) SFT + PPO: reinforcement learning with Proximal Policy Optimization applied on top of SFT; (3) SFT + GRPO: process-level preference alignment achieved by applying Group Relative Policy Optimization following SFT.

All three models were trained and evaluated under identical data conditions and inference settings to ensure fairness and comparability. The experimental results are summarized in Table 2.

From an overall perspective, the three training strategies exhibit a clear progressive improvement. First, relying solely on SFT already brings a substantial performance gain, increasing the overall accuracy from the base model's 67.50% to 77.40%. This demonstrates that high-quality, structured agricultural reasoning-chain data plays a crucial role in establishing fundamental diagnostic ability, effectively enhancing the model's performance across multiple tasks such as disease identification, pest judgment, and maturity assessment. Second, introducing PPO on top of SFT further improves the reasoning process. With PPO, the model's accuracy increases to 79.76%, indicating that policy optimization helps the model generate more stable reasoning chains under uncertain conditions. Among the three strategies, SFT + GRPO achieves the best overall performance, reaching an accuracy of 80.67% and leading across the three major domains of Plant Disease, Pest Infestation, and Maturity. Compared with PPO, the advantage of GRPO lies primarily in its more robust policy-constraint mechanism, which prevents drastic fluctuations in reasoning behavior

Table 2
Comparison of different training strategies on Lychee-Bench.

Model	Plant Disease	Pest Infestation	Maturity	Overall
Qwen2.5VL-7B	69.6	66.2	65.5	67.50
Qwen2.5VL-7B+SFT	79.0	78.2	71.4	77.40
Qwen2.5VL-7B+SFT+PPO	81.2	79.7	76.3	79.76
Qwen2.5VL-7B+SFT+GRPO	83.0	80.2	76.0	80.67

Table 3
Comparison of model performance under two data formats using the same training strategy (SFT + GRPO).

Dataset	Model	Plant Disease	Pest Infestation	Maturity	Overall
Flat QA	Qwen2.5VL-7B	69.6	66.2	65.5	67.50
	Qwen2.5VL-7B+SFT	77.3	75.5	69.4	72.44
	Qwen2.5VL-7B+SFT+GRPO	83.4	78.7	72.2	79.58
QA+CoT	Qwen2.5VL-7B	69.6	66.2	65.5	67.50
	Qwen2.5VL-7B+SFT	79.0	78.2	71.4	77.40
	Qwen2.5VL-7B+SFT+GRPO	83.0	80.2	76.0	80.67

during updates. This allows the model to retain the domain-knowledge structure learned during SFT while further strengthening its ability to aggregate key visual evidence. Observations of the model's reasoning traces show that GRPO produces more coherent chains, enabling more accurate step-by-step verification of disease symptoms, pest morphology, or maturity indicators, and performing effective correction when contradictions arise—an ability that is particularly crucial for agricultural diagnostic tasks involving multi-factor coupling and cross-symptom reasoning.

In summary, all training strategies contribute to performance improvement, but the GRPO-based two-stage training framework achieves the best balance among accuracy, reasoning stability, and evidence integration capability. This demonstrates its effectiveness and suitability for high-risk agricultural diagnostic scenarios.

3.4.2. Ablation on data formats and reasoning structures

In agricultural visual diagnostic tasks, the model must handle more than surface-level visual classification; it must engage in comprehensive cognitive reasoning involving multi-factor coupling, cross-stage inference, and causal relationship identification. Therefore, whether the training data explicitly presents the reasoning chain is crucial for determining whether the model can learn and reliably reproduce the diagnostic processes used by agricultural experts. Based on this understanding, this section aims to systematically analyze how different data formats contribute to shaping the model's reasoning capability.

To isolate the independent effect of data structure, we construct two mutually exclusive data formats: (1) the Flat QA dataset, which contains only images, questions, and final answers, without any intermediate reasoning steps, simulating the conventional supervised learning paradigm; and (2) the QA+CoT dataset, which provides a complete "dynamic trial-and-error trajectory," including initial hypothesis formation, visual evidence verification, contradiction detection, and conclusion refinement. This structured formulation is designed to explicitly teach the model the logical pathways underlying agricultural diagnosis. In the experimental setup, the base model architecture and training strategy remain identical, with the only variable being the training data format. As shown in Table 3, the performance gap during the supervised fine-tuning stage is the most pronounced. When trained with Flat QA, the model's overall accuracy increases from 67.50% to 72.44%, yielding only a modest improvement. In contrast, using the QA+CoT format raises accuracy from 67.50% to 77.40%, an increase of 9.9%, nearly double the gain achieved with Flat QA. This substantial difference demonstrates that structured data providing explicit reasoning chains leads to significantly higher learning efficiency in agricultural diagnostic tasks, enabling the model to establish a more stable reasoning foundation at an early training stage.

Furthermore, this advantage persists during the reinforcement learning stage. Under the same SFT + GRPO training strategy, the model trained with QA+CoT achieves an overall accuracy of 80.67%, notably higher than the

79.58% obtained with Flat QA. This indicates that structured reasoning chains not only enhance the model's foundational capabilities during supervised learning but also improve the effectiveness of policy optimization when reward signals are introduced, leading to more stable and coherent reasoning trajectories.

Based on the experimental results, the advantages of the QA+CoT data format can be further explained from three complementary perspectives. (1) The structured reasoning chain provides a more solid cognitive foundation, enabling the model to grasp essential diagnostic reasoning steps—such as hypothesis generation, evidence verification, and conflict identification—early in the learning process. (2) QA+CoT improves the consistency and interpretability of intermediate reasoning steps, allowing the model to generate more coherent reasoning trajectories even under challenging conditions such as multi-symptom coupling, noisy backgrounds, or complex field environments. (3) The explicit reasoning process enhances the model's ability to utilize reward signals during reinforcement learning, enabling GRPO to more effectively identify high-value reasoning behaviors and perform stable optimization, thereby further improving the model's performance in high-difficulty diagnostic tasks.

In summary, the data format plays a fundamental role in training agricultural visual-reasoning models. The structured QA+CoT format substantially improves the model's reasoning quality, stability, and reliability, thereby establishing a critical foundation for practical deployment in high-risk agricultural production scenarios.

4. Discussion

In this study, a multimodal cognitive reasoning framework (Lychee-Mind) tailored for agricultural scenarios was constructed based on Lychee-Bench, and systematic evaluations were conducted across twelve categories of fine-grained tasks. Experimental results demonstrate that the framework maintains high robustness and consistency under real field images, complex environmental conditions, and cross-task reasoning requirements, thereby highlighting the effectiveness of process-level supervision and the two-stage training paradigm in agricultural visual reasoning tasks. When compared with various mainstream open-source and closed-source multimodal models under the same parameter scale, Lychee-Mind exhibits significant superiority, particularly achieving remarkable advantages in high-complexity tasks such as Anomaly Reasoning (AR), Climatic and Environmental Reasoning (CTR), and Agronomic Planning and Decision-Making (APD). Collectively, these findings support a core insight: in the highly open and uncertain vertical domain of agriculture, improving the structural quality of model reasoning processes is more valuable than merely optimizing final answers.

Firstly, the findings of this study underscore the significance of "process alignment" in agricultural artificial intelligence. Traditional methods rely on final answers for supervision, namely "outcome alignment", This approach is susceptible to data biases in open environments, leading to insufficient stability in the construction of model reasoning chains and difficulty in handling atypical or long-tailed samples. In contrast, the "process reward function based on evidence aggregation" designed in this study explicitly shifts the optimization objective from "generating correct answers" to "conducting professional and consistent reasoning." This encourages the model to proactively perform hypothesis establishment, evidence verification, and contradiction identification during the reasoning process, thereby aligning more closely with the diagnostic logic of agricultural experts. In experiments, the two-stage training strategy performed particularly prominently in high-difficulty tasks, indicating that process constraints play a crucial role in multi-factor coupled inference within agricultural scenarios.

Secondly, high-quality and structured domain-specific data played a critical role in this study. The Lychee-Dataset was constructed based on 6,553 images collected from real agricultural fields, encompassing a total of 29,756 annotated samples. Through explicit Chain-of-Thought (CoT) annotations, the diagnostic process was refined from "directly providing conclusions" into a series of step-by-step observation and reasoning procedures. Such structured CoT annotations not only significantly enhanced the models behavioral imitation capability during the Supervised Fine-Tuning (SFT) phase but also provided quantifiable and evaluable reasoning bases for the design of process rewards in the reinforcement learning stage. In ablation experiments, the performance improvement achieved by structured CoT was significantly higher than that of traditional Flat QA data, indicating that data structure density is far more critical than data volume itself in vertical agricultural domains. Agricultural images are characterized by high noise, variable conditions, and severe long-tailed distributions; therefore, relying solely on final labels often makes it difficult for models to grasp general reasoning logic, a limitation that can be effectively addressed by structured reasoning chains.

Thirdly, the Lychee-Mind framework demonstrates favorable scalability and generalization potential. Although this study focuses on lychee as a cash crop, its core methodologysimulating experts self-correction and evidence-driven reasoning processes through two-stage trainingpossesses cross-crop universality. The same mechanism can be transferred to diagnostic tasks for other high-value cash crops such as citrus, grapes, and mangoes, and even extended to other agricultural sub-fields including livestock husbandry, aquaculture, and greenhouse crop management. New interpretable agricultural agents can be constructed by replacing domain-specific datasets and adjusting the evaluation criteria of the reward function according to task characteristics. Such transferability holds potential value for the establishment of regional agricultural technology service systems and the enhancement of rural digital infrastructure.

Despite achieving promising results, this study still has several limitations. Firstly, the current evaluation benchmark, consisting of 500 questions, is relatively limited in scale and has not covered more challenging complex scenarios such as concurrent multiple diseases, severe environmental interference, or cross-regional differences, which may lead to insufficient comprehensive evaluation of the model. Secondly, the process reward in this study relies on large closed-source models such as GPT-4o to assess the quality of reasoning chains, which introduces additional computational costs and potential domain biases to a certain extent. Future research needs to explore lightweight reward models with stronger domain adaptability, or reduce reliance on external evaluation models through distillation and self-supervised learning methods. Additionally, the current framework has not fully utilized multi-source data such as meteorological data, soil sensor data, and time-series images; in the future, multi-modal information fusion will be considered to enhance the models stability across multiple scenarios and time scales.

Overall, based on structured reasoning chains, process rewards, and a two-stage training paradigm, this study provides a novel technical pathway for agricultural multimodal agents. In the future, the reliability, interpretability, and deployability of the model under real agricultural production conditions will be further enhanced by expanding data scenarios, optimizing reward mechanisms, and introducing multi-source fusion technologies.

5. Conclusion

In this work, we present Lychee-Mind, a domain-specialized multimodal framework for lychee pest and disease diagnosis as well as maturity assessment. By constructing Lychee-Dataset, a large-scale, process-level annotated dataset, designing a two-stage training paradigm, and establishing Lychee-Bench, a fine-grained reasoning benchmark, we systematically assess the effectiveness and necessity of combining structured chain-of-thought (CoT) supervision with process-level reinforcement learning (RL) in complex agricultural scenarios. Experimental results show that, compared with the base model, Lychee-Mind improves overall accuracy on Lychee-Bench from 67.50% to 80.67%. For the three core tasks of disease diagnosis, insect pest identification, and maturity assessment, accuracy increases from 69.6%, 66.2%, and 65.5% to 83.0%, 80.2%, and 76.0%, respectively, substantially enhancing diagnostic robustness and task generalization under real field conditions. From the perspective of the training paradigm, the two-stage pipeline clearly reveals the marginal contribution of process-level rewards. Supervised fine-tuning (SFT) alone raises overall accuracy to 77.40%. On top of this, policy optimization with PPO further improves accuracy to 79.76%, while the evidence-aggregation-based GRPO scheme pushes performance to 80.67%. These results indicate that a reward mechanism centered on the quality of the reasoning process can effectively suppress shortcut or logically inconsistent behaviors, strengthen causal consistency, and enable the model to maintain stable and trustworthy reasoning under multiple confounding factors and complex visual backgrounds. Comparisons with 24 mainstream large multimodal models with comparable parameter counts further show that Lychee-Mind achieves leading overall performance and excels on several high-difficulty tasks, demonstrating favorable accuracyparameter efficiency. Analysis of the generated reasoning behavior shows that Lychee-Mind produces structured and interpretable CoT reasoning in challenging cases such as weakly textured lesions, highly cluttered backgrounds, ambiguous symptom presentation, and cross-symptom inference. Such capabilities are particularly important for high-risk agricultural decision-making, as they facilitate human inspection of intermediate reasoning and enhance agronomy experts trust in model outputs. This, in turn, provides a practical basis for embedding multimodal agents into real-world production workflows, including pest and disease early warning, field scouting, and precision pesticide application (Lan et al., 2017). Overall, this study demonstrates the substantial potential of process-oriented multimodal training paradigms for agricultural cognitive intelligence, and verifies the synergistic effect of high-quality domain data, structured CoT supervision, and process-level reward design in improving the trustworthiness, interpretability, and stability of agricultural AI systems. Future work will extend the proposed framework to additional high-value crops and regional scenarios, explore lightweight model distillation and multimodal temporal reasoning, and investigate more autonomous process reward learning strategies, thereby further promoting the large-scale deployment of cognitive agricultural agents in real production environments.

CRediT authorship contribution statement

Guanyu Zhu: Writing original draft, Methodology, Conceptualization, Software, Investigation & editing. **Kunming Lv:** Writing. **Hanbing Liu:** Formal analysis & review. **Fengqi Yu:** Methodology, Writing original draft & Data curation. **LinHu:** Visualization. **Ersheng Ni:** Writing. **Meiyi Lu**:Visualization. **Weixing Wang:** Funding acquisition & Writing review**Jiaxing Xie:** Funding acquisition & Writing review & editing.

Declaration of Competing Interest:

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

Acknowledgments: This work was supported by the Guangzhou Key Research and Development Program (2025B03J0100), the China Agriculture Research System of MOF and MARA(CARS-32-10), the Guangdong Modern Agriculture Industry Technology System Innovation Team Project(2024CXTD19-11) and the College Students' Innovative Entrepreneurial Training Plan Program,China(202510564040),

References

Food and Agriculture Organization of the United Nations (FAO). World food and agriculture statistical yearbook 2023. Technical report, Food and Agriculture Organization of the United Nations, 2023. URL https://doi.org/10.4060/cc8166en.

P. Chen et al. Field study of uav variable-rate spraying method for orchards based on canopy volume. *Agriculture*, 15(13):1374, 2024a. doi: 10.3390/agriculture15131374. URL https://doi.org/10.3390/agriculture15131374.

Y. Lan et al. Time series model for predicting the disturbance of lychee canopy by wind field in unmanned aerial spraying system. *Computers and Electronics in Agriculture*, 2025. doi: 10.1016/j.compag.2025.109954. URL https://doi.org/10.1016/j.compag.2025.109954.

C. Shengde et al. Detection and tracking of agricultural spray droplets using gsconv-enhanced yolov5s and deepsort. *Computers and Electronics in Agriculture*, 2025. doi: 10.1016/j.compag.2025.110353. URL https://doi.org/10.1016/j.compag.2025.110353.

S. Li et al. A multi-scale cucumber disease detection method in natural scenes based on yolov5. *Computers and Electronics in Agriculture*, 2022. doi: 10.1016/j.compag.2022.107363. URL https://doi.org/10.1016/j.compag.2022.107363.

S. Yin et al. Foundation models in agriculture: A comprehensive review. *Agriculture*, 15(8):847, 2025. doi: 10.3390/agriculture15080847. URL https://doi.org/10.3390/agriculture15080847.

C. Wang and Z. Xiao. Lychee surface defect detection based on deep convolutional neural networks with gan-based data augmentation. *Agronomy*, 11(8):1500, 2021. doi: 10.3390/agronomy11081500. URL https://doi.org/10.3390/agronomy11081500.

J. Xie, X. Zhang, Z. Liu, F. Liao, W. Wang, and J. Li. Detection of litchi leaf diseases and insect pests based on improved fcos. *Agronomy*, 13(5): 1314, 2023. doi: 10.3390/agronomy13051314. URL https://doi.org/10.3390/agronomy13051314.

J. Xiao, G. Kang, L. Wang, Y. Lin, F. Zeng, J. Zheng, R. Zhang, and X. Yue. Real-time lightweight detection of lychee diseases with enhanced yolov7 and edge computing. *Agronomy*, 13(12):2866, 2023. doi: 10.3390/agronomy13122866. URL https://doi.org/10.3390/agronomy13122866.

C.-Y. Wang et al. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2023. doi: 10.1109/CVPR52729.2023.00721. URL https://doi.org/10.1109/CVPR52729.2023.00721.

J. Wang et al. A top-view transformer-based framework for tiny tea bud detection with gradient-guided attention and position-sensitive loss. *Computers and Electronics in Agriculture*, 2025. doi: 10.1016/j.compag.2025.111123. URL https://doi.org/10.1016/j.compag.2025.111123.

C. Zhou et al. From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models. 2025. URL https://arxiv.org/abs/2509.25373.

Q. Li et al. Can large multimodal models understand agricultural scenes? benchmarking with agromind. 2025. doi: 10.48550/arXiv.2505.12207. URL https://doi.org/10.48550/arXiv.2505.12207.

L. Fang et al. Agri-llm: Prompt-based large language model for emission data analytics in smart agriculture. *IEEE Transactions on Industrial Informatics*, 2024. doi: 10.1109/TII.2024.3411942. URL https://ieeexplore.ieee.org/document/11194249.

B. Yang et al. Agrigpt: A large language model ecosystem for agriculture. 2025. URL https://arxiv.org/abs/2508.08632.

L. Wang et al. Agri-llava: Knowledge-infused large multimodal assistant on agricultural pests and diseases. 2024. doi: 10.48550/arXiv.2412.02158. URL https://doi.org/10.48550/arXiv.2412.02158.

A. Gauba et al. Agmmu: A comprehensive agricultural multimodal understanding benchmark. 2025. doi: 10.48550/arXiv.2504.10568. URL https://doi.org/10.48550/arXiv.2504.10568.

M. Awais et al. Agrogpt: Efficient agricultural vision-language model with expert tuning. 2024. doi: 10.48550/arXiv.2410.08405. URL https://doi.org/10.48550/arXiv.2410.08405.

J. Zhou et al. Agri-bench: A hierarchical agriculture benchmark. 2024. doi: 10.48550/arXiv.2403.09876. URL https://doi.org/10.48550/arXiv.2403.09876.

W. Huang et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. 2025. doi: 10.48550/arXiv.2503.06749. URL https://doi.org/10.48550/arXiv.2503.06749.

OpenAI. Gpt-4o system card. Technical report, OpenAI, 2024. URL https://openai.com/research/gpt-4o-system-card.

G.D. Team Gemini. Gemini 1.5: Unlocking multimodal understanding across modalities. 2024. URL https://arxiv.org/abs/2403.05530.

Team Gemini. Gemini 1.5 flash model card. Technical report, Google, 2024. URL https://storage.googleapis.com/deepmind-media/gemini/gemini-1.5-flash-model-card.pdf.

H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024a. doi: 10.1109/CVPR52729.2024.02629. URL https://arxiv.org/abs/2310.03744.

H. Liu, C. Li, and LLaVA Team. Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-02-29-llava-next/, 2024b. Accessed: 2025-11-22.

Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. 2024b. URL https://arxiv.org/abs/2312.14238.

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, and J. Tang. Qwen2.5-vl technical report. 2025. doi: 10.48550/arXiv.2502.13923. URL https://arxiv.org/abs/2502.13923.

01-AI. Yi: Open foundation models by 01.ai. 2024. doi: 10.48550/arXiv.2403.05530. URL https://arxiv.org/abs/2403.05530.

Y. Yao, J. Yan, Y. Zhang, et al. Minicpm-v: A gpt-4v-level multimodal model on your phone. 2024. doi: 10.48550/arXiv.2408.01800. URL https://arxiv.org/abs/2408.01800.

M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, et al. Phi-3 technical report: A highly capable language model locally on your phone. 2024. doi: 10.48550/arXiv.2404.14219. URL https://arxiv.org/abs/2404.14219.

H. Liu, C. Li, and LLaVA Team. Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024c. Accessed: 2025-11-23.

Y. Lan, S. Chen, and B. K. Fritz. Current status and future trends of precision agricultural aviation technologies. *Precision Agriculture*, 2017. doi: 10.1007/s11119-017-9523-4. URL https://api.semanticscholar.org/CorpusID:114585944.

Highlights

- Constructed the Lychee dataset, a large-scale multimodal collection featuring structured Chain-of-Thought (CoT) annotations to enable precise orchard diagnosis.
- Developed Lychee-Mind, a domain-specialized large multimodal model tailored for lychee pest, disease, and maturity assessment.
- Proposed a two-stage training paradigm combining supervised fine-tuning with GRPO-based reinforcement learning and a process-level evidence reward.
- Established Lychee-Bench, a comprehensive benchmark covering 12 fine-grained tasks to systematically evaluate agricultural visual reasoning capabilities.