

Thinking Like a Botanist: Challenging Multimodal Language Models with Intent Driven Chain-of-Inquiry

Anonymous ACL submission

Abstract

Vision evaluations are typically done through multi-step processes. In most contemporary fields, experts analyze images using structured, evidence-based adaptive questioning. In plant pathology, botanists inspect leaf images, identify visual cues, infer diagnostic intent, and probe further with targeted questions that adapt to species, symptoms, and severity. This structured probing is crucial for accurate disease diagnosis and treatment formulation. Yet current vision-language models are evaluated on single-turn question answering. To address this gap, we introduce PlantInquiryVQA, a benchmark for studying multi-step, intent-driven visual reasoning in botanical diagnosis. We formalize a Chain of Inquiry framework modeling diagnostic trajectories as ordered question-answer sequences conditioned on grounded visual cues and explicit epistemic intent. We release a dataset of 24,964 expert-curated plant images and 138,078 question-answer pairs annotated with visual grounding, severity labels, and domain-specific reasoning templates. Evaluations on top-tier Multimodal Large Language Models reveal that while they describe visual symptoms adequately, they struggle with safe clinical reasoning and accurate diagnosis. Importantly, structured question-guided inquiry significantly improves diagnostic correctness, reduces hallucination, and increases reasoning efficiency. We hope PlantInquiryVQA serves as a foundational benchmark in advancing research to train diagnostic agents to reason like expert botanists rather than static classifiers.

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015) datasets have become a central paradigm for evaluating multimodal reasoning, with applications spanning medical imaging, scientific image analysis, and embodied agents (Das et al., 2017; Abacha et al., 2019; Liu et al., 2024). Recent VQA benchmarks Wei et al. (2022) have largely con-

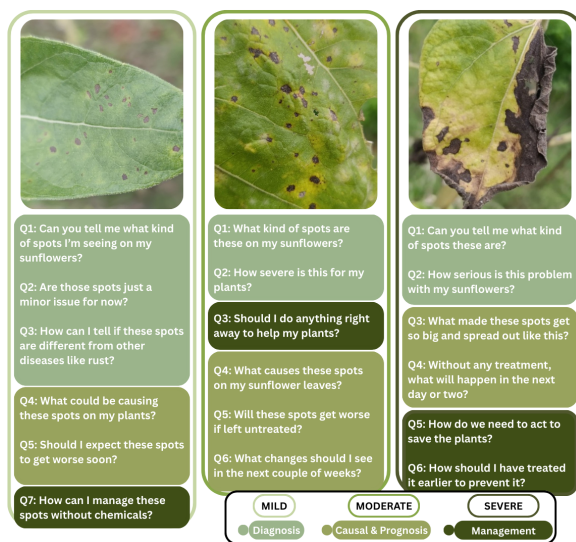


Figure 1: Unlike static QA datasets that ask generic questions regardless of disease status, our framework aligns the epistemic intent of the inquiry with visual severity. The diagram illustrates how the questioning focus evolves with disease progression in Sunflower (Alternaria Leaf Spot). The cognitive task transitions from Ambiguity Resolution (Diagnosis) in early stages to Future Forecasting (Prognosis) and Action Planning (Management) in advanced stages, ensuring questions are contextually relevant to the visual evidence.

tributed in improving Vision Language Models' capability of understanding complex multiobject-multifocus scenarios. In **VQA**, images are analyzed in the context of a given question, requiring strong understanding of both visual cues and natural language processing (Antol et al., 2015; Andreas et al., 2016; Abacha et al., 2019). Advanced **VQA** datasets now focus on multipanel, multichoice and strong visual-language grounded question-answer (QA) pairs (Fan et al., 2024; Liu et al., 2024). These large datasets enable **Multimodal Large Language Models (MLLMs)** to have more nuanced understanding and generation of content that blends both visual and linguistic elements. Despite these advances, most existing VQA

059 benchmarks and plant-focused vision datasets re- 111
 060 main fundamentally question-centric: they treat 112
 061 each image as an independent input to a single 113
 062 query or a static set of QA pairs, rather than as 114
 063 the starting point of a goal-directed, adaptive in- 115
 064 quiry. In agricultural vision specifically, widely 116
 065 used datasets primarily target classification and 117
 066 segmentation tasks, such as plant disease recog- 118
 067 nition and leaf-level diagnosis, and do not capture 119
 068 the hierarchical, evidence-conditioned questioning 120
 069 strategies employed by domain experts (Singh et al., 121
 070 2020; Mohanty et al., 2016). 122

071 However, in many real-world applications, effective 123
 072 visual reasoning does not arise from answering 124
 073 isolated questions. Instead, it emerges from a 125
 074 deliberate sequence of interdependent inquiries, 126
 075 where each question is conditioned on prior obser- 127
 076 vations, and follows a sequential narrative trajec- 128
 077 tory (Pearl and Mackenzie, 2018; Andreas et al., 129
 078 2016). The sequence and intent of the questions 130
 079 are as critical as the answers themselves. This is 131
 080 particularly evident in the field of botanical sci- 132
 081 ence where each plant sample is given unique con- 133
 082 sideration based on its visual appearance. Expert 134
 083 botanists conduct the holistic evaluation of a leaf 135
 084 sample, from species identification, to disease diag- 136
 085 nosis and prognosis-prediction through a process of 137
 086 structured hierarchical and evidence-driven ques- 138
 087 tioning strategy (Agrios, 2005; Schumann et al., 139
 088 1991). This process, which we refer to as a **Chain- 140
 089 of-Inquiry** (CoI), is grounded in visual cues identi- 141
 090 fied from the sample image and varies substantially 142
 091 depending on the plant’s health condition. For ex- 143
 092 ample, when disease symptoms are ambiguous, ex- 144
 093 perts prioritize differential diagnosis, comparative 145
 094 visual analysis, and disease progression prediction 146
 095 (Strange, 2003). Conversely, for samples exhibiting 147
 096 severe damage and clearly manifested symptoms, 148
 097 the expert CoI shifts toward disease management, 149
 098 prevention strategies, and earlier intervention sce- 150
 099 narios. As a result, each plant sample is treated as 151
 100 a distinct case, warranting its own adaptive ques- 152
 101 tioning strategy. 153

102 Though Large Language Models have made 154
 103 significant progress in implementing **Chain-of- 155
 104 Thought** (CoT) reasoning for high level planning 156
 105 and task completion, similar multistep exploration 157
 106 remains underexplored in designing VQA datasets 158
 107 (Wei et al., 2022; Kojima et al., 2022). This gap 159
 108 raises a fundamental question: how well can simi- 160
 109 lar structured probing be translated through visual 161
 110 Question Answer pairs. In this paper we introduce

PlantInquiryVQA a multi-step hierarchical ques- 111
 tion answering benchmark with 24,964 curated and 112
 expert validated plant images and 138k correspond- 113
 ing question answer pairs related to the botanical 114
 field. We design 7 distinct question categories capa- 115
 ble of extracting all relevant information regarding 116
 a plant sample’s holistic evaluation. We then clas- 117
 sify 12 unique visual-cue dependent and narrative 118
 driven adaptive *CoI*. Our implementation strategy 119
 allows for each plant sample to have its unique 120
 questioning trajectory specific to its visual cues 121
 with sequence of interconnected questions and an- 122
 swers. These serve as the precondition for in-depth 123
 multistage reasoning and contextual semantic anal- 124
 ysis. Finally, we benchmark popular open-sourced 125
 and proprietary MLLMs on the PlantInquiryVQA 126
 benchmark and conduct thorough error analysis 127
 with the help of expert verification. 128

Our contribution includes: 129

1. **The PlantInquiryVQA Benchmark:** We re- 130
 lease a large-scale dataset of 25k manually 131
 curated images across diverse crop species, 132
 annotated with expert-verified visual cue de- 133
 scriptions and domain-specific knowledge 134
 bases. 135
2. **The Chain-of-Inquiry (CoI) Framework:** 136
 We formalise a novel reasoning taxonomy for 137
 PlantInquiryVQA, classifying 12 unique reason- 138
 ing templates into 7 distinct cognitive cat- 139
 egories (including Etiological Reasoning, Dif- 140
 ferential Diagnosis, and Counterfactual Anal- 141
 ysis). 142
3. **Diagnostic Reasoning Evaluation:** We con- 143
 duct a comprehensive evaluation of both 144
 closed- and open-source MLLMs. Our results 145
 show that question-guided protocols signifi- 146
 cantly reduce hallucination and improve diag- 147
 nostic correctness, while sequential chains of 148
 inquiry enhance explainability efficiency. 149

2 Related Works 150

Visual Question Answering (VQA). Antol et al. 151
 (2015) combines image understanding with natural 152
 language processing to answer queries about visual 153
 content. Unlike standard classification, VQA fa- 154
 cilitates interactive question–answering, allowing 155
 models to capture and articulate complex relation- 156
 ships within an image. As such, the application 157
 of VQA now spans multiple domains. These in- 158
 clude: educational tools, customer service systems, 159

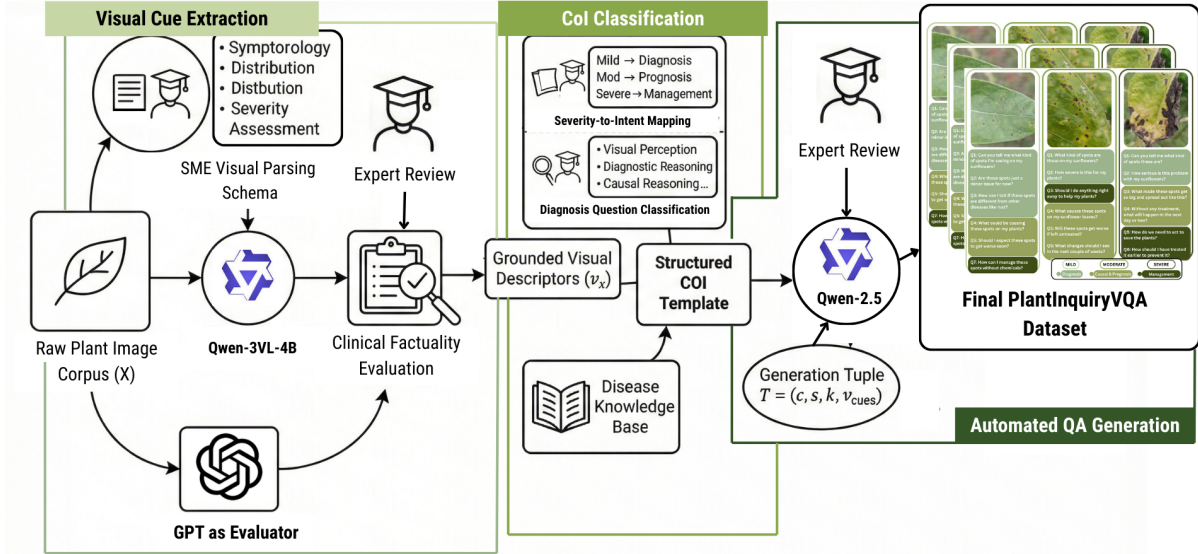


Figure 2: Overall Methodology Pipeline for PlantInquiryVQA CoI Dataset Generation. The process is divided into three phases: (1) Extracting grounded visual cues using VLM guided by expert schemas; (2) Structuring botanical knowledge to map disease severity to diagnostic intent; and (3) A dynamic LLM generation pipeline that injects specific reasoning modules based on the determined intent and visual evidence.

and autonomous driving etc. In particular, VQA shows exceptional potential in the field of pathological diagnosis and health inquiry (Zhang et al., 2023a). Current medical VQA benchmarks include PMC-VQA Zhang et al. (2023a), SLAKE Liu et al. (2021), Path-VQA He et al. (2020), and VQA-RAD Lau et al. (2018). However, these datasets are focused on medical diagnostics. In agriculture, existing popular datasets like PlantVillage (Hughes and Salathé, 2015), PlantDoc Singh et al. (2020), and PlantSeg Wolny et al. (2020) focus on classification or segmentation tasks. While they support disease detection, they do not enable interactive reasoning through question–answer formats.

Chain-of-Thought in Multimodal Models. Parallel to these advances, benchmark datasets have also evolved to test deeper cognitive capabilities through Chain-of-Thought (CoT). Originally proposed for text-based Large Language Models by Wei et al. (2022), CoT marked a paradigm shift by prompting models to decompose complex problems into intermediate logical steps rather than mapping inputs directly to outputs. This methodology has recently been adapted for Multimodal Large Language Models (MLLMs), where the reasoning chain must ground linguistic tokens in visual features to reduce hallucination (Zhang et al., 2023b). Recent frameworks like Multimodal-CoT Zhang et al. (2023b); Liu et al. (2024) demonstrate that incorporating explicit reasoning paths signif-

icantly improves performance on complex vision-language tasks. However, in the majority of these works, CoT is treated primarily as a prompting strategy or a latent capability of the model architecture, rather than an explicit structural requirement inherent to the dataset itself.

Structured and Hierarchical VQA. Consequently, there have been efforts to translate this step-by-step reasoning into VQA dataset design. Recent benchmarks have attempted to introduce structure into visual questioning; for instance, BloomVQA Gong et al. (2024) organizes questions based on Bloom’s taxonomy of cognitive complexity, while MedCoT Liu et al. (2024) utilizes hierarchical expert agents to simulate medical reasoning flows. Similarly, conversational benchmarks like Das et al. (2017) introduced the concept of multi-turn visual dialogue. However, most existing hierarchical datasets rely on static question taxonomies or crowd-sourced dialogues that lack the goal-oriented precision of a domain expert (Gong et al., 2024). They generally fail to capture the causal dependency of professional diagnosis, where the formulation of the next question is strictly conditional on the visual evidence verified in the previous step. **PlantInquiryVQA** fills this gap by making the CoI, an explicit dataset-level artifact: the dataset’s question sequences are constructed to mirror the adaptive, decision-driven workflows of domain experts.

3 Methodology

3.1 Formalization of Chain-of-Inquiry

We define the **CoI** as a visual-semantic trajectory conditioned on diagnostic intent. Let \mathcal{X} denote the set of plant images and \mathcal{V} be the space of explicit visual cues (e.g., *chlorosis patterns*, *lesion margins*). For a given image $x \in \mathcal{X}$, we extract a set of grounded visual descriptors $v_x \in \mathcal{V}$.

CoI C is defined as an ordered sequence of T dialogue turns:

$$C(x, v_x) = \langle (q_1, a_1), (q_2, a_2), \dots, (q_T, a_T) \rangle \quad (1)$$

where each question q_t is conditioned on the visual evidence v_x , the previous context $H_{t-1} = \{(q_i, a_i)\}_{i=1}^{t-1}$, and a latent diagnostic intent $k \in \mathcal{K}$. Here, the intent space \mathcal{K} is stratified into three primary epistemic goals derived from expert botany:

1. **Diagnosis (k_D):** Identifying health status and discriminating between similar pathologies (**Differential Diagnosis**).

2. **Prognosis (k_P):** Predicting disease trajectory, temporal evolution, and causal etiology.

3. **Management (k_M):** Prescribing remediation strategies and formulating counterfactual preventative reasoning.

Thus, the generation of a specific CoI is modeled as sampling from a conditional distribution $P(C | x, v_x, k)$, ensuring that the dialogue trajectory aligns with the clinical necessity of the plant sample.

3.2 Visual Cue Extraction and CoI Classification

In extracting relevant visual cues from a sample image, we strictly adhere to the **Symptomatological Diagnostic Protocol**, as defined in [Agrios \(2005\)](#). We recruited two PhD-level and four graduate-level botanists specializing in plant pathology as Subject Matter Expert (SMEs) to define a "Visual Parsing Schema". Following the diagnostic criteria outlined by [Agrios \(2005\)](#); [Streets \(1972\)](#), the experts established a structured feature extraction template comprising three critical diagnostic dimensions: *Symptomatology*, *Distribution Patterns*, *Disease Severity Quantification*. A comprehensive description of these criteria are provided in the appendix [A.5](#). In the pilot phase, each SME applied this schema to a randomized batch of 50 images. To ensure inter-annotator consistency, batches were cross-verified among the group, synthesizing a unified, expert-validated template.

Automated Extraction & Validation. Leveraging this expert-derived schema, we prompted three open-source Vision-Language Models (VLMs) to generate dense, fine-grained visual cue descriptions for a batch of 500 images. As shown by the comparative benchmarking in [Table 1](#), **Qwen3-VL-4B** outperformed all other models. Consequently, we used **Qwen3-VL-4B** to extract visual cues for the entire corpus of 24,964 images. Given the scale of the dataset, exhaustive expert annotation was infeasible. We therefore employed a hybrid validation strategy: (1) Qwen3-VL extracted visual cues using expert-designed schemas; (2) **GPT-4V Achiam et al. (2023)** acted as an external evaluator to cross-verify the extracted cues, flagging outputs with high semantic divergence from Qwen3-VL; (3) finally, domain experts conducted a rigorous **Clinical Factuality** check on all flagged instances plus 5000 randomly sampled images. Experts annotated errors in two categories: *Object Hallucination* (mentioning symptoms, e.g., "halo", not present in the image) and *Attribute Mismatch* (incorrectly describing color or texture). The model achieved a Factuality Score of 93.8%, defined as the proportion of generations free from critical clinical errors. [Appendix A.5.3](#) includes the comparative analysis of model-extracted visual cues together with the expert-annotated, knowledge-based cues.

We rely on human specialists to collect high-quality groundtruth QA chains for plant pathology diagnosis. We ask our botanists to compile relevant and semantically rich question samples from established botanical sources ([Agrios, 2005](#); [Schumann et al., 1991](#); [Strange, 2003](#); [Streets, 1972](#)). While classical literature outlines the biological phases of diagnosis (e.g., Symptomatology, Etiology, Epidemiology), there is no standard taxonomy for interrogating them in a visual dialogue. To bridge this gap, our SMEs classify the standard diagnostic inquiries in 7 general categories: **Visual Perception & Grounding:** Corresponds to the *Symptomatology* ([Strange, 2003](#)). **Diagnostic Reasoning:** Aligns with the *Differential Diagnosis* ([Streets, 1972](#)). **Causal Reasoning:** Derived from *Etiology*. **Risk Assessment:** Maps to *Epidemiology*. **Prognosis Prediction:** Models the *Disease Cycle*. **Prescriptive Reasoning:** Corresponds to *Disease Management*. **Counterfactual Reasoning:** Simulates retrospective analysis (e.g., 'What if treatment had occurred earlier?'), a crucial component of post-epidemic evaluation and learning.

We choose a random sample of 600 images from

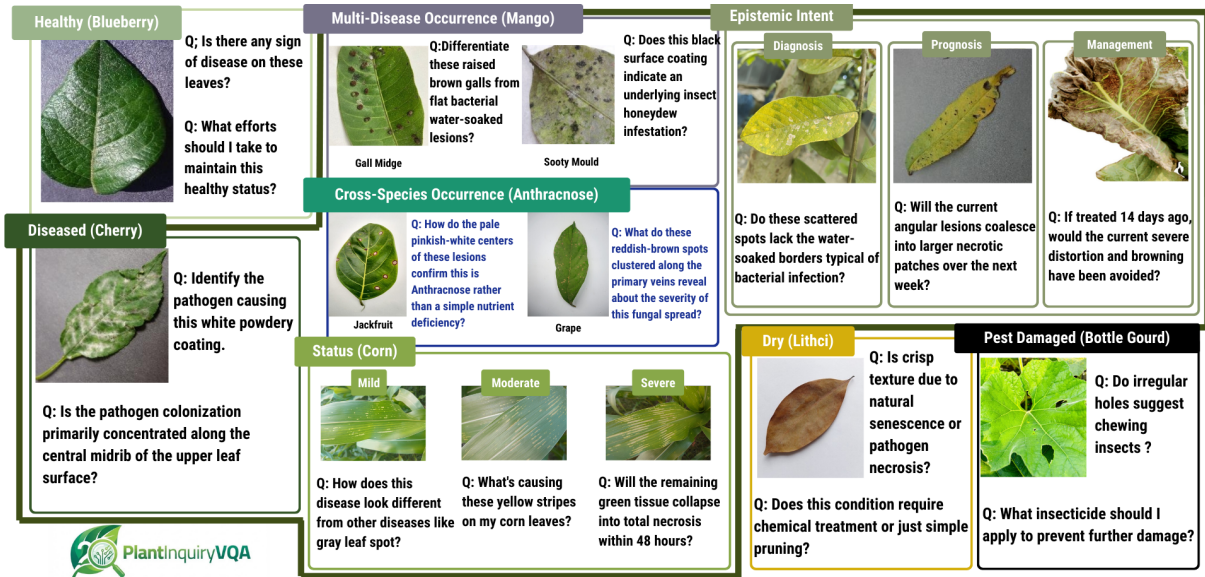


Figure 3: Qualitative Examples of 12 Distinct CoI Trajectories. The framework adapts questioning strategies across four varying axes of diagnostic complexity: (1) Health Status (Healthy, Diseased, Senescence, Pest Damaged); (2) Disease Severity, showing evolving logic for Mild, Moderate, and Severe Maize infection; (3) Instance Variety, handling Multi-disease and Cross-species constraints; and (4) Epistemic Intent, illustrating the cognitive shift from Diagnosis to Prognosis and Management based on visual evidence.

our dataset and ask the experienced botanists to conduct a clinical evaluation of each sample and note down their questioning strategy. Here we identify three distinct intent evident among the questioning arc for a given plant image. These intents are explained in 3.1 and further explained here: For samples exhibiting **Mild** or ambiguous symptoms, the CoI intent **Diagnosis** (k_D) focuses on *differential diagnosis*. Botanists compare disease symptoms and ask clarifying questions to distinguish the observed pathology from potential look-alikes. Conversely, for samples classified as **Moderate**, the focus shifts to **Causal & Prognosis** (k_P). Here, the inquiries emphasize understanding the disease’s spread and inferring the environmental conditions such as humidity or soil pH that likely facilitated the infection. Finally, for **Severe** cases where damage is extensive, the expert intent shifts to **Management** (k_M). The reasoning chain here is primarily focused on identifying immediate strategies for disease remedy and employing counterfactual analysis (e.g., "How would the plant’s condition differ if intervention had occurred two weeks prior?") to simulate critical care scenarios. Following this schema, We translate the botanist chain of thought into 12 distinct **Chain-of-Inquiry**, covering 4 Health Status (Healthy, Diseased, Senescence, Pest Damaged), 3 Disease Severity (Mild, Moderate, Severe), 2 Instance Variety (Multi-disease Oc-

Model	Score (%)
Qwen3-VL	184 (73.6%)
Gemma3-4B	58 (23.2%)
LLaVA-1.5	8 (3.2%)

Model	Score (%)
Qwen3-VL	185 (74.0%)
Gemma3-4B	61 (24.4%)
LLaVA-1.5	4 (1.6%)

Table 1: Comparative analysis of three open-source models for visual grounding (top) and specificity (bottom).

currence, Cross-species Occurrence) and 3 Epistemic Intent (Diagnosis, Prognosis, Management). Appendix A.7 shows the each CoI with corresponding image and varying question samples.

3.3 Dataset Generation Pipeline

We developed a structured generation pipeline governed by the configuration tuple $T = (c, s, k_s, V_{cues})$, corresponding to *Biological condition, Severity, Severity derived Intent, and Visual cues*. This decoupling allows us to generate diverse reasoning chains even for the same image (e.g., asking for *Management* advice on a *Mild* case vs. a *Severe* case). Here, we employed Qwen2.5-7B-Instruct (Yang et al., 2025) to dynamically assemble dialogue trajectories from question tem-

plates. Here, the epistemic goal k ($Diagnosis(k_D)$, $Prognosis(k_P)$ and $Management(k_M)$) modulates the information density based on severity status s . To further enhance complexity, we inject specific reasoning modules (e.g., `temporal_evolution`, `remediation_strategy` e.t.c) directly into the chains. Figure 2 shows the overall generation of the **PlantInquiryVQA** dataset. Our tiered approach ensures that PlantInquiryVQA covers the full spectrum of diagnostic difficulty, from routine identification to complex, multi-step clinical reasoning. We provide the detailed module injection logic and generation protocols in Appendix A.6.

4 Experiment

We first evaluate eighteen leading Multimodal Large Language Models (MLLMs) on the PlantInquiry VQA benchmark. Then, based on the evaluation results, we conduct a thorough error analysis and ablation study to isolate the impact of our CoI framework on diagnostic reasoning.

4.1 Experimental Setup

MLLMs The MLLMs we adopt in the evaluation include both open-weight models and proprietary models accessed via API. The open-weight MLLMs include the **Qwen** series (Qwen3-VL-32B, Qwen2.5-VL-72B/32B, Qwen-VL-Plus), **Gemma** variants (Gemma-3-27B), and the **Ministral** family (Ministral-8B/3B). We also evaluate a distilled lightweight model, **Nemotron-Nano-12B**, to assess performance at the edge. For proprietary models, we evaluate **Gemini-2.5-Pro**, **Gemini-2.5-Flash**, and the recently released **Gemini-3-Flash**.

Evaluation In our evaluation process, we employ a dual-metric strategy. For linguistic quality, we report standard lexical metrics (F1, BLEU-4, ROUGE-L). However, given the clinical nature of the task, lexical overlap is often insufficient. Therefore, we introduce domain-specific alignment scores: **Clinical Factuality**, **Disease Accuracy**, and **Visual Grounding Quality (VG)**. These metrics are defined in appendix A.1.

4.2 Main Result

We evaluate the dataset using a Cumulative Context Test, in which each successive question is conditioned on the full history of preceding questions and generated answers in the chain. Comparative performance across models is reported in Table 2. The results reveal that **Gemini-3-Flash**

model consistently outperforms all other evaluated architectures, establishing a highest on this benchmark. It achieves the highest scores across both standard lexical metrics and domain specific alignment score. This performance gap suggests that advanced closed-source models possess superior instruction-following capabilities, allowing them to better handle the long context dependencies required by CoI. Unlike smaller models that often lose track of the diagnostic "narrative" after several turns, Gemini-3-Flash maintains coherence, accurately translating visual symptoms into precise disease identifications and safety-compliant advice.

Interestingly, while Gemini-3-Flash excels in clinical reasoning, the **GroK-4.1-Fast** maintains the highest Visual Grounding (VG) score. This indicates that while models these model excels at "accurately describing raw visual features such as "yellow spots" or "necrotic margins, they struggle to synthesize that evidence into a coherent clinical diagnosis. This discrepancy highlights that "seeing" the symptom is not equivalent to "diagnosing" the pathology, underscoring the necessity of our **CoI** framework to bridge this cognitive lapse.

The most prominent finding from Table 2 is the substantial domain gap in current state-of-the-art MLLMs regarding botanical pathology. Even the top-performing model, Gemini-3-Flash, achieves a Clinical Utility score of only 0.188 and a Disease Identification score of 0.444. This indicates that PlantInquiry VQA represents a significantly hard benchmark; models struggle to translate visual signals into accurate, safe clinical diagnoses.

4.3 Error Analysis

Intending to identify why models fail or succeed within our framework, we move beyond simple accuracy metrics. We leverage the structured nature of PlantInquiry VQA to conduct comparative experiments that isolate specific reasoning capabilities.

How does the structure of inquiry influence diagnostic accuracy? To evaluate if the questions themselves act as effective attention mechanisms, we conducted a **Protocol Structure Benefit Test**. We compared model performance under two conditions: (1) *Direct Diagnosis*, where the model is simply asked to identify the disease from the image.

Model	Lexical Metrics			Domain Alignment & Quality				
	F1	BLEU-4	R-L	Dis	Clin.	Safe.	VG	Len.
GEMINI-3-FLASH	0.255	0.033	0.196	0.444	0.188	0.147	0.259	85.8
GEMINI-2.5-PRO	0.225	0.016	0.132	0.357	0.112	0.040	0.408	142.9
QWEN3-VL-235B	0.210	0.013	0.120	0.348	0.111	0.035	0.489	143.9
SEED-1.6-FLASH	0.226	0.022	0.139	0.344	0.120	0.075	0.394	99.1
LLAMA-3.2-90B-VISION	0.212	0.014	0.105	0.340	0.185	0.214	0.372	134.9
LLAMA-4-MAVERICK	0.212	0.013	0.103	0.329	0.175	0.202	0.397	144.5
GEMINI-2.5-FLASH	0.226	0.018	0.145	0.299	0.098	0.046	0.392	163.5
QWEN3-VL-32B	0.182	0.011	0.096	0.288	0.096	0.035	0.475	227.8
GEMMA-3-27B	0.192	0.011	0.103	0.272	0.086	0.032	0.353	156.9
PIXTRAL-12B	0.225	0.016	0.122	0.272	0.145	0.159	0.368	98.0
QWEN2.5-VL-32B	0.177	0.009	0.076	0.254	0.078	0.017	0.463	260.4
PHI-4-MULTIMODAL	0.177	0.010	0.097	0.254	0.087	0.040	0.358	167.2
QWEN2.5-VL-72B	0.236	0.016	0.123	0.247	0.080	0.040	0.375	106.2
GROK-4.1-FAST	0.203	0.016	0.132	0.224	0.067	0.009	0.498	100.7
MISTRAL-MEDIUM-3.1	0.211	0.015	0.119	0.205	0.062	0.023	0.360	110.7
MINISTRAL-8B	0.180	0.010	0.094	0.197	0.060	0.020	0.394	151.8
MINISTRAL-3B	0.166	0.007	0.083	0.189	0.059	0.020	0.372	163.0

Table 2: **Main Results on Test Set.** GEMINI-3-FLASH demonstrates significant superiority, achieving the highest scores across all lexical and domain-specific semantic metrics. SEED-1.6-FLASH shows competitive performance with strong disease accuracy (0.344). LLAMA-3.2-90B-VISION and LLAMA-4-MAVERICK exhibit high clinical utility and safety scores. QWEN-VL-PLUS maintains the highest Visual Grounding (VG) quality (0.508), followed by GROK-4.1-FAST (0.498), suggesting a trade-off between clinical reasoning accuracy and pure descriptive grounding.

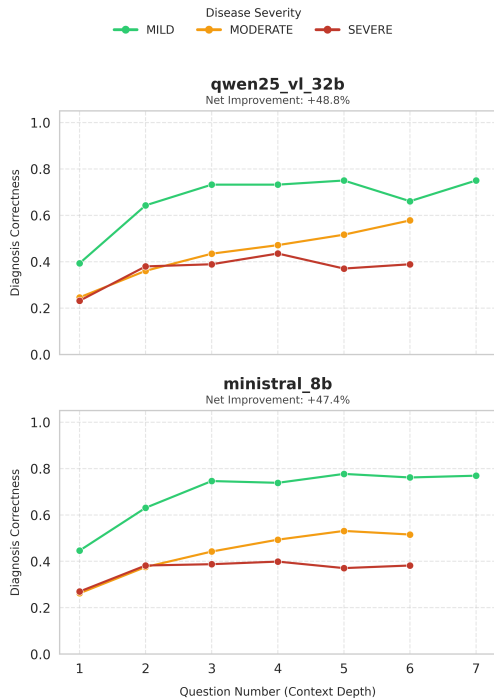


Figure 4: Protocol Structure Benefit Test for Qwen25-VL-32B and Ministral-8B. Both models achieve better Diagnosis Correctness scores using questions as context (48.8% and 47.4%, respectively).

(2). *Question-Guided*, where the model is provided with the list of diagnostic questions (e.g., “Are the margins water-soaked?”) before mak-

ing a diagnosis. From Figure 4, we observe that the *Question-Guided* condition yields significantly higher **diagnosis Correctness** across all three disease status (*mild*, *moderate* and *severe*) compared to the direct approach. In the Direct condition, models often hallucinate common diseases based on prior bias (e.g., assuming “Early Blight” for any tomato leaf). However, the specific questions force the model to attend to fine-grained features (lesion margins or halo presence), effectively constraining the search space and reducing hallucination. Appendix A.8 shows the accuracy trajectory as the inquiry progresses for all other models.

Does the Chain of Inquiry promote reasoning efficiency? A key hypothesis of our work is that structured inquiry should lead to more efficient information retrieval, rather than just “chattier” responses. To test this, we conducted a *Ratio Test* to measure the **Explainability Efficiency** (E), defined as the ratio of verified visual cues extracted per 100 words generated (see Appendix A.1 for E definition). We compared two settings: *Scaffolded*, the model answers questions in isolation without access to the previous dialogue history and *Guided*, the model answers questions sequentially, with the ground-truth history of previous turns provided as context.

Model	Scaffolded			Guided			$\Delta\%$
	Eff.	Cues	Words	Eff.	Cues	Words	
GROK-4.1-FAST*	4.54	8.49	220.6	5.20	5.80	135.0	14.54
GEMINI-2.5-FLASH	2.60	8.65	456.0	3.67	4.71	181.0	41.15
GEMINI-2.5-PRO	2.95	6.11	268.6	3.58	4.00	147.9	21.36
QWEN3-VL-32B	2.88	8.82	362.6	3.33	5.96	206.4	15.63
SEED-1.6-FLASH	3.22	5.63	198.1	3.75	4.10	125.0	16.46
QWEN2.5-VL-32B	1.60	6.81	471.4	2.94	4.47	249.2	83.75
QWEN2.5-VL-72B	2.46	4.53	212.0	2.92	3.47	171.9	18.70
NEMOTRON-NANO-12B	3.84	5.72	164.7	3.34	4.32	165.4	(-14.97)
PIXTRAL-12B	2.53	4.00	217.1	2.90	3.20	145.0	14.62
LLAMA-3.2-90B-VISION	2.40	4.18	197.6	2.85	3.50	140.0	18.75
GEMMA-3-27B	1.88	6.61	411.0	2.38	4.77	256.2	26.60
MINISTRAL-3B	2.26	5.30	247.8	2.71	3.30	152.1	19.91
MINISTRAL-8B	2.21	5.19	264.2	2.65	3.70	172.4	19.91
QWEN-VL-PLUS	1.63	7.85	536.5	2.53	4.92	282.0	55.21
MISTRAL-MEDIUM-3.1	2.35	4.75	228.6	2.70	3.60	155.0	14.89
LLAMA-4-MAVERICK	2.31	4.32	208.7	2.65	3.40	145.0	14.72
PHI-4-MULTIMODAL	1.94	6.71	423.4	2.55	4.80	240.0	31.44

Table 3: **Visual Feature Extraction Efficiency.** Green cells denote the higher-performing prompting strategy per model. $\Delta\%$ indicates relative improvement of the better setting over the weaker one. *Eff.* = Efficiency Score, *Cues* = Average Cues Found, *Words* = Average Word Count.

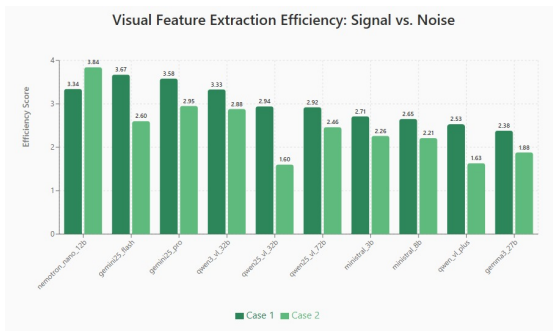


Figure 5: Ratio Test comparison across Scaffolded (case 1) and Guided (case 2) settings. It can be seen that the **Explainability Efficiency** increases for all models excluding NEMOTRON-NANO-12B.

The results, depicted in Table 3, reveal a crucial insight into how models utilize context: **1.** Capable models like Gemini-2.5-Flash and Qwen3-VL show a significant increase in efficiency in the Guided setting (e.g., Gemini-2.5-Flash improves from 2.60 to 3.67). When provided with the chain history, these models stop “hedging” or repeating basic observations (e.g., “This is a leaf”) and focus entirely on the new, specific visual evidence required by the current question. **2.** The lightweight Nemotron-Nano-12B experienced a performance drop in the Guided setting (3.84 to 3.34). This suggests that smaller models may struggle with “context distraction,” where a longer dialogue history interferes with, rather than aids, the immediate visual processing task.

5 Conclusion

In this work, we introduced **PlantInquiryVQA**, a benchmark designed to shift agricultural AI from static classification to dynamic, expert-level diagnostic reasoning. By formalizing the *Chain-of-Inquiry* framework, we modeled the adaptive questioning strategies of botanists, creating a dataset of 24,964 curated images and 138k dialogue turns that simulate real-world clinical workflows. Our extensive evaluation reveals a critical reality: while current MLLMs demonstrate impressive visual perception, a significant “domain gap” remains in their ability to perform safe clinical reasoning. Although Gemini-3-Flash achieved state-of-the-art performance, its Clinical Utility score (0.188) and Safety score (0.147) indicate that models are not yet reliable enough for autonomous deployment. Furthermore, we identified a distinct trade-off between visual grounding and logical deduction, with models like Grok-4.1-Fast often describing symptoms more accurately than they diagnose them. However, our ablation studies confirm that the CoI structure itself, specifically the use of guided, intent-driven inquiry, significantly enhances reasoning efficiency, reducing hallucination and increasing the information density of model responses. We hope PlantInquiryVQA serves as a foundational testbed for the next generation of “AI botanists”, agents capable not just of seeing a disease, but of reasoning through its management to secure global food systems.

538 Limitations

539 While PlantInquiryVQA represents a significant
540 advancement in agricultural visual reasoning, sev-
541 eral limitations constrain its immediate applicabil-
542 ity. Primarily, real-world plant pathology requires
543 holistic sensory inputs beyond a single static im-
544 age, such as tactile feedback (texture) and environ-
545 mental context (soil moisture), which our single-
546 frame optical dataset cannot replicate. Furthermore,
547 despite the structured CoI framework, our exper-
548 iments show that even top-tier models continue
549 to hallucinate safety, occasionally classifying dis-
550 eased samples as healthy; this "false reassurance"
551 poses a tangible risk, necessitating that this sys-
552 tem be used currently as a decision-support tool for
553 experts rather than an autonomous replacement. Fi-
554 nally, the current benchmark is exclusively English-
555 based, which limits accessibility for smallholder
556 farmers in non-English speaking regions who stand
557 to benefit most from automated diagnostics.

558 References

559 Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey
560 Liu, Dina Demner-Fushman, and Henning Müller.
561 2019. VQA-Med: Overview of the medical visual
562 question answering task at ImageCLEF 2019. In
563 *CLEF 2019 Working Notes*.

564 Joshua Achiam, Steven Adler, Sandhini Agarwal, Lama
565 Ahmad, Ilge Akkaya, Florencia Aleman, Diogo
566 Almeida, Janko Altmenschmidt, Sam Altman, Shub-
567 ham Anadkat, and 1 others. 2023. Gpt-4 technical
568 report. *arXiv preprint arXiv:2303.08774*.

569 George N Agrios. 2005. *Plant pathology*. Elsevier.

570 Kouser Ahamed, Rokonozzaman Ayon, Mst. Momotaz
571 Begum, and Israt Jahan. 2025. *Bdlitchi: A field-*
572 *collected bangladeshi litchi leaf disease dataset for*
573 *deep learning-based detection and classification*.

574 Md Hasan Ahmad. 2024. *Advanced tea crop disease*
575 *study: High-resolution dataset for precision agricul-*
576 *ture and pathological insight*.

577 Faruk Ahmed and Ferdous Ahmed. 2025. *Tld-bd: A*
578 *comprehensive tea leaf image dataset for leaf con-*
579 *dition analysis*.

580 Tanvir Ahmed, Mst Taposi Rabeya Taposi,
581 Marzanul Alam Mukhor, and Mayen Uddin
582 Mojumdar. 2024. *Plant leaf disease recognition*
583 *dataset*.

584 Hasan Al Banna, Md. Fuad Hossain, and Mayen Uddin
585 Mojumdar. 2024. *Papaya leaf disease image dataset*.

S. Ali, M. Ibrahim, S. I. Ahmed, M. Nadim, M. Rahman,
M. M. Shejunti, and T. Jabid. 2022. *MangoLeafBD:*
A comprehensive image dataset to classify mango
leaf diseases. 586
587
588
589

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and
Dan Klein. 2016. Neural module networks. In *Pro-*
ceedings of the IEEE conference on computer vision
and pattern recognition, pages 39–48. 590
591
592
593

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Mar-
garet Mitchell, Dhruv Batra, C Lawrence Zitnick, and
Devi Parikh. 2015. Vqa: Visual question answering.
In *Proceedings of the IEEE international conference*
on computer vision, pages 2425–2433. 594
595
596
597
598

Leo Prasanth Lourdu Antony. 2023. *Rice leaf diseases*
dataset. 599
600

Rokonozzaman Ayon, Nur Yea Afroz Suchi,
Md Asif Shahriar Arpon, Tanjina Ahmed Tuly, and
Abdullah Al Noman. 2025. *Arabian jasmine leaf*
condition dataset (ajlcd-2025). 601
602
603
604

Puja Bapari, Md Zinnahtur Rahman Zitu, and Mst Um-
mehani. 2025. *A high-resolution image dataset of*
tomato (solanum lycopersicum) leaves for multi-class
disease detection and classification from bangladesh. 605
606
607
608

Swami Nisha Bhagirath, Vaibhav Bhatnagar, and Linesh
Raja. 2023. *Winter wheat leaf images dataset*. 609
610

Md Masum Billah, Saifuddin Sagor, and Moham-
mad Shorif Uddin. 2025a. *Hibiscus leaf diseases*
classification dataset. 611
612
613

Md Masum Billah, Saifuddin Sagor, and Moham-
mad Shorif Uddin. 2025b. *A real-world hibiscus*
and tea leaf image dataset for classification. 614
615
616

Prayma Bishshash, Md Asraful Sharker Nirob,
Md Habibur Shikder, and Afjal Sarower. 2024. *Sar-*
cld-2024: A comprehensive dataset for cotton leaf
disease detection. 617
618
619
620

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh,
Deshraj Yadav, José MF Moura, Devi Parikh, and
Dhruv Batra. 2017. Visual dialog. In *Proceedings of*
the IEEE conference on computer vision and pattern
recognition, pages 326–335. 621
622
623
624
625

Utsab Das, Showrov Azam, and Md Abdullah Al Kafi.
2025. *Banana and banana leaf dataset for classifica-*
tion and disease detection. 626
627
628

Pulak Deb Nath. 2025. *Citrusleafvision: A diverse*
dataset for lemon leaf disease detection. 629
630

Pulak Deb Nath, Faruk Ahmed, and Belal Uddin. 2025.
Bdrubberleaf: A comprehensive dataset of rubber
tree leaf diseases from bangladesh for agricultural
research. 631
632
633
634

Emerson M Del Ponte, Sarah J Pethybridge, Clive H
Bock, Sami J Michereff, Franklin J Machado, and
Piérri Spolti. 2017. Standard area diagrams for aiding
severity estimation: scientometrics, pathosystems, 635
636
637
638

639	and methodological trends in the last 25 years. <i>Phytopathology</i> , 107(10):1161–1174.	Image dataset for turmeric plant leaf disease detection.	693
640			694
641	Madhuri Dharrao, Deepak Dharrao, Rakesh Sonawane, and Nilima Zade. 2025. Niphad grape leaf disease dataset (ngld) .	Shahariar Hossain, Saifuddin Sagor, and Faruk Ahmed. 2025c. High-resolution images of lychee plant diseases for classification and detection .	695
642			696
643			697
644	Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Yang Zhao, Xinze Guan, and Xin Wang. 2024. Muffin or chihuahua? challenging multimodal large language models with multipanel vqa. In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 6845–6863.	Md Faysal Hossain and Md Zamirul Islam Reyad. 2025. Comprehensive mango leaf images dataset for multi-class disease classification and automated plant disease detection .	698
645			699
646			700
647			701
648			702
649			703
650			704
651	Yunye Gong, Robik Shrestha, Jared Claypoole, Michael Cogswell, Arijit Ray, Christopher Kanan, and Ajay Divakaran. 2024. Bloomvqa: Assessing hierarchical multi-modal comprehension. In <i>Findings of the Association for Computational Linguistics: ACL 2024</i> , pages 14905–14918.	Rashidul Hasan Hridoy, Md Tarek Habib, Imran Mahmud, Aminul Haque, and Md Abdulla Al Mamun. 2025. Black gram leaf image dataset for disease detection in field conditions .	705
652			706
653			707
654			708
655			709
656			710
657	Md Ehsanul Haque and Md Al-Imran. 2025. Agrileafnet: Fruit tree leaf dataset for agricultural research .	David P. Hughes and Marcel Salathé. 2015. An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing . <i>CoRR</i> , abs/1511.08060.	711
658			712
659			713
660	Md. Taufiq Hasan, Sabbir Hossain Durjoy, Md Emon Shikder, Md. Safaet Zahangir, Md Muntasir Uddin, Md Mehedi Hasan Shoib, and Md Hasan Imam Bijoy. 2025a. Litchileaf4001: A comprehensive dataset of lychee leaf diseases for ai-based visual diagnosis .	Rezwan Huq, Farzia Hossain, Shahida Begum, Raiyan Gani, and Jubaer Ahmed. 2025. Image datasets for jackfruit plant leaf disease .	714
661			715
662			716
663			717
664			718
665	Md. Zahid Hasan, Abu Kowshir Bitto, and Md Hasan Imam Bijoy. 2025b. Brinjalfruitx: A field-collected image dataset for machine learning and deep learning-based disease identification in brinjal fruits .	ICAR-CITH. 2024. Apple leaf diseases image dataset of ICAR-CITH .	719
666			720
667			721
668			722
669			723
670	Mehedi Hasan, Sonia Khatun, Md Abu Raihan, and Abdul Hasib Uddin. 2023. Rice leaf bacterial and fungal disease dataset .	Ahmed Imtiaz, Fahad Bin Islam Swapnil, Syed Rayhan Masud, and Debajoyti Karmaker. 2024. Tomato leaf dataset: A dataset for multiclass disease detection and classification .	724
671			725
672			726
673	Rakib Hasan, Sanimun Hossain Sanzit, Md Mamun Hosen, Ferdous Hasan, Md Mehedi Hasan Topu, and Muksitul Islam. 2025c. High-resolution eggplant leaf image dataset for plant disease classification and detection .	Md Forhadul Islam, Imon Sutradhar, and Md Mizanur Rahman. 2025. Leaf image dataset for disease detection in bitter gourd, okra, pumpkin, and ridge gourd .	727
674			728
675			729
676			730
677			731
678	Sadib Hassan, Md Zahid Akon, Partho Sarathi Sarker, Rahat Hossain, Jannatul Ferdous, Md Mahadi Hasan Shaon, and Md Abdur Razzak. 2025. Ugv: Guava leaves disease dataset bangladesh .	Abid Khan, Sabit Ahamed Preanto, Tapon Paul, and Md Hasan Imam Bijoy. 2025. Moringaleafnet: A multi-class leaf disease dataset for precision agriculture and deep learning research .	732
679			733
680			734
681			735
682	Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. <i>arXiv preprint arXiv:2003.10286</i> .	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. <i>Advances in neural information processing systems</i> , 35:22199–22213.	736
683			737
684			738
685			739
686	Ismail Hossain, Khandaker Rezoanul Haque, Abdullah Al Rafi, and Md Zahidul Islam Talukder. 2025a. Tomato leaf diseases .	Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In <i>2021 IEEE 18th international symposium on biomedical imaging (ISBI)</i> , pages 1650–1654. IEEE.	740
687			741
688			742
689	Md Riyad Hossain, Mohammad Rifat Ahmmad Rashid, Tasfia Binte Jahangir, Md Samir Hossain, Md Mahamudur Rahman, Raiyan Gani, Jubaer Ahmed, Raihan Ul Islam, and M Saddam Hossain Khan. 2025b.	Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. <i>Scientific data</i> , 5(1):1–10.	743
690			744
691			745
692			746
		Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024. Medcot: Medical chain of thought via hierarchical expert. <i>arXiv preprint arXiv:2412.13736</i> .	

746	Yongbo Liu. 2025. Tomato disease dataset .	Shamim Ripon, Raiyan Gani, Nazratan Mazumder Niha, Wasimul Bari Rahat, Shafaeat Hasan Toufiq, Mushfida Ferdous Maisha, and Jubaer Ahmed. 2025. Cotton leaf image dataset for disease classification .	796 797 798 799
747 748	Laurence V Madden, Gareth Hughes, and F van den Bosch. 2007. <i>The study of plant disease epidemics</i> .	Saifuddin Sagor, Md Faysal Hossan, Faruk Ahmed, and Md Zamirul Islam Reyad. 2025. Sunflower plant health and growth stage image dataset for agricultural machine learning applications .	800 801 802 803
749 750	Eram Mahamud and Md Assaduzzaman Tapos. 2024. Lentil plant disease image dataset (4 class) .	Gail Lynn Schumann and 1 others. 1991. <i>Plant diseases: their biology and social impact</i> . APS Press, American Phytopathological Society St. Paul, MN, USA.	804 805 806 807
751 752 753	Md Mafiul Hasan Matin, Mehedi Hasan Hasan, and Sabik Ur Rahman. 2025. Cauliflower diseases identification image dataset .	Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. 2020. Plantdoc: A dataset for visual plant disease detection. In <i>Proceedings of the 7th ACM IKDD CoDS and 25th COMAD</i> , pages 249–253.	808 809 810 811 812
754 755 756 757	Sharada P Mohanty, David P Hughes, and Marcel Salathé. 2016. Using deep learning for image-based plant disease detection. <i>Frontiers in plant science</i> , 7:215232.	Richard N Strange. 2003. <i>Introduction to plant pathology</i> . John Wiley & Sons.	813 814
758 759 760 761 762	Md Asraful Sharker Nirob, Prayma Bishshash, Mariyam Bin Ayan, Tania Khatun, and Mohammad Shorif Uddin. 2024. Eggplant dataset: A comprehensive dataset for agricultural research and disease detection .	Rubert B Streets. 1972. <i>The diagnosis of plant diseases: a field and laboratory manual emphasizing the most practical methods for rapid identification</i> . University of Arizona Press.	815 816 817 818
763 764 765 766	Md Asraful Sharker Nirob, Prayma Bishshash, A K M Fazlul Kobir Siam, and Mayen Uddin Mojumdar. 2025a. Cair-bgd-2025: Annotated dataset for bottle gourd disease & growth stages .	Nusrat Sultana, Sumaita Binte Shorif, Morium Akter, and Mohammad Shorif Uddin. 2022. Cucumber disease recognition dataset .	819 820 821
767 768 769 770	Md Asraful Sharker Nirob, A K M Fazlul Kobir Siam, Prayma Bishshash, and Md Assaduzzaman. 2025b. Chili plant leaf disease and growth stage dataset from bangladesh .	Ranjita Thapa, Kai Zhang, Noah Snaveley, Serge Belongie, and Awais Khan. 2020. The plant pathology challenge 2020 data set to classify foliar disease of apples . <i>Applications in Plant Sciences</i> , 8(9):e11390.	822 823 824 825
771 772 773	Judea Pearl and Dana Mackenzie. 2018. <i>The book of why: the new science of cause and effect</i> . Basic books.	Sandip Thite and Kailas Patil. 2025. Pisum sativum image dataset: Healthy and disease-affected cases .	826 827
774 775 776	Md Istiak Rahman Radowan and Rokonozzaman Ayon. 2025. Disease dataset of wheat: Original, augmented, and balanced for deep learning .	Nguyen Thanh Truong, Xuan Linh Nguyen, Pham Dinh Cap Thang, and Tuong Le. 2025. A durian leaf image dataset of common diseases in vietnam for agricultural diagnosis .	828 829 830 831
777 778 779	Maruful Islam Rafe, Farhan Masud Nayem, Shanto Babu Sarker, and Abdullah Al Shiam. 2025. Eggplant_leaf_disease_dataset .	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. <i>Advances in neural information processing systems</i> , 35:24824–24837.	832 833 834 835 836 837
780 781 782 783	Salman Af Rahman, Md Nafiz Imtiaz, Naima Ahmed, and Md Hasan Imam Bijoy. 2025. Burmese grape leaf disease dataset for computer vision-based plant health diagnosis .	Adrian Wolny, Lorenzo Cerrone, Athul Vijayan, Rachele Tofanelli, Amaya Vilches Barro, Marion Louveaux, Christian Wenzl, Sören Strauss, David Wilson-Sánchez, Rena Lymbouridou, and 1 others. 2020. Accurate and versatile 3d segmentation of plant tissues at cellular resolution. <i>Elife</i> , 9:e57613.	838 839 840 841 842 843
784 785 786 787 788	Aditya Rajbongshi, Umme Sara, Bonna Akter, Rashiduzzaman Shakil, and Sadia Sazzad. 2022. Sunflower fruits and leaves dataset for sunflower disease classification through machine learning and deep learning .	An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, and 1 others. 2025. Qwen2. 5-1m technical report. <i>arXiv preprint arXiv:2501.15383</i> .	844 845 846 847 848
789 790 791 792	Mohammad Rifat Ahmmad Rashid, Raiyan Gani, Jubaer Ahmed, Maherun Nessa Isty, and Sawkat Ali. 2024. Healthy and unhealthy papaya leaf images from bangladeshi orchards .		
793 794 795	Shakhawath Hossain Rifat, Tanvir Almas Layes, Afif Hasan, and Mayen Uddin Mojumdar. 2024. Rice leaf disease and pest dataset overview .		

849 Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weix-
850 iong Lin, Ya Zhang, Yanfeng Wang, and Weidi
851 Xie. 2023a. Pmc-vqa: Visual instruction tuning for
852 medical visual question answering. *arXiv preprint*
853 *arXiv:2305.10415*.

854 Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao,
855 George Karypis, and Alex Smola. 2023b. Multi-
856 modal chain-of-thought reasoning in language mod-
857 els. *arXiv preprint arXiv:2302.00923*.

858 Md Zinnahtur Rahman Zitu, Shahariar Rahman Shifat,
859 and Mayen Uddin Mojumdar. 2024. [A benchmark](#)
860 [dataset for detecting disease in plant leaves: An es-](#)
861 [sential resource for deep learning models.](#)

A Appendix 862

This appendix provides additional details and re- 863
resources referenced in the main paper. 864

A.1 Plant Disease Datasets 865

section A provides a comprehensive overview of 866
publicly available plant disease datasets used in 867
PlantInquiryVQA and their licensing information. 868

Dataset Name	License	Ref.
PlantVillage	CC0	(Hughes and Salathé, 2015)
Chili Plant Leaf Disease	CC BY 4.0	(Nirob et al., 2025b)
Banana and Banana Leaf Dataset	CC BY 4.0	(Das et al., 2025)
Bottle Gourd Disease & Growth Stages	CC BY 4.0	(Nirob et al., 2025a)
Plant Leaf Disease Recognition Database	CC BY 4.0	(Ahmed et al., 2024)
Papaya Leaf Disease	CC BY 4.0	(Al Banna et al., 2024)
Eggplant Leaf Disease Dataset	CC BY 4.0	(Rafe et al., 2025)
Lychee Plant Diseases	CC BY 4.0	(Hossain et al., 2025c)
TLD-BD (Tea Leaf)	CC BY 4.0	(Ahmed and Ahmed, 2025)
Leaf Disease (Bitter Gourd, Okra, Pumpkin)	CC BY 4.0	(Islam et al., 2025)
Advanced Tea Crop Disease Study	CC BY 4.0	(Ahmad, 2024)
Plant Pathology Challenge 2020 (Apple)	CC BY 4.0	(Thapa et al., 2020)
Lentil Plant Disease	CC BY 4.0	(Mahamud and Tapos, 2024)
SAR-CLD-2024 (Cotton)	CC BY 4.0	(Bishshash et al., 2024)
Apple Leaf Diseases (ICAR-CITH)	CC BY 4.0	(ICAR-CITH, 2024)
Hibiscus and Tea Leaf	CC BY 4.0	(Billah et al., 2025b)

869

Dataset Name	License	Ref.	Dataset Name	License	Ref.
MoringaLeafNet	CC BY 4.0	(Khan et al., 2025)	Comprehensive Mango Leaf	CC BY 4.0	(Hossan and Reyad, 2025)
BDLitchi	CC BY 4.0	(Ahamed et al., 2025)	AgriLeafNet	CC BY 4.0	(Haque and Al-Imran, 2025)
Mango leaf datasets	CC BY 4.0	(Ali et al., 2022; Hossan and Reyad, 2025)	Pisum sativum (Pea)	CC BY 4.0	(Thite and Patil, 2025)
Sunflower Fruits and Leaves	CC BY 4.0	(Rajbongshi et al., 2022)	Turmeric Plant Leaf Disease	CC BY 4.0	(Hossain et al., 2025b)
Multifaceted Rose Leaf Disease	CC BY 4.0	(?)	Niphad Grape Leaf Disease (NGLD)	CC BY 4.0	(Dharrao et al., 2025)
Real-World Papaya Leaf	CC BY 4.0	(Rashid et al., 2024)	LitchiLeaf4001	CC BY 4.0	(Hasan et al., 2025a)
Hibiscus Leaf Diseases Classification	CC BY 4.0	(Billah et al., 2025a)	Eggplant leaves (multiple datasets)	CC BY 4.0	(Nirob et al., 2024; Rafe et al., 2025)
Tomato disease datasets	CC BY 4.0	(Liu, 2025; Imtiaz et al., 2024; Bapari et al., 2025)	Sunflower Plant Health & Growth Stage	CC BY 4.0	(Sagor et al., 2025)
Arabian Jasmine (AJLCD-2025)	CC BY 4.0	(Ayon et al., 2025)	Burmese Grape Leaf Disease	CC BY 4.0	(Rahman et al., 2025)
Durian leaf diseases (Vietnam)	CC BY 4.0	(Truong et al., 2025)	Black Gram Leaf	CC BY 4.0	(Hridoy et al., 2025)
BrinjalFruitX	CC BY 4.0	(Hasan et al., 2025b)	Tomato Leaf Diseases (additional)	CC BY 4.0	(Hossain et al., 2025a; Imtiaz et al., 2024; Liu, 2025)
BDRubberLeaf	CC BY 4.0	(Deb Nath et al., 2025)	UGV Guava Leaves Disease (Bangladesh)	CC BY 4.0	(Hassan et al., 2025)
Cauliflower Diseases	CC BY 4.0	(Matin et al., 2025)	CAIR-BGD-2025 (Bottle Gourd)	CC BY 4.0	(Nirob et al., 2025a)
CitrusLeafVision (Lemon)	CC BY 4.0	(Deb Nath, 2025)	High-Resolution Eggplant Leaf	CC BY 4.0	(Hasan et al., 2025c)
Tomato (Bangladesh high resolution)	CC BY 4.0	(Bapari et al., 2025)	Banana Leaves Imagery Dataset	CC BY 4.0	(Das et al., 2025)
Jackfruit Plant Leaf Disease	CC BY 4.0	(Huq et al., 2025)	Rice Leaf Bacterial and Fungal Disease	CC BY 4.0	(Hasan et al., 2023)
Cotton Leaf Image Dataset	CC BY 4.0	(Bishshash et al., 2024; Ripon et al., 2025)	Rice Leaf Disease and Pest Overview	CC BY 4.0	(Rifat et al., 2024)

Dataset Name	License	Ref.
Rice Leaf Diseases Dataset	CC BY 4.0	(Antony, 2023)
Disease Dataset of Wheat	CC BY 4.0	(Radowan and Ayon, 2025)
Winter Wheat Leaf Images	CC BY-NC 3.0	(Bhagirath et al., 2023)
Cucumber Disease Recognition	CC BY 4.0	(Sultana et al., 2022)
Benchmark Dataset for Plant Leaves	CC BY 4.0	(Zitu et al., 2024)

This datasets in total contain 321k plant images across 39 crop species and 119 total disease.

A.1 Semantic and Clinical Metrics

To assess the model’s reliability in a high-stakes agricultural context, we define four domain-specific metrics beyond standard lexical overlap. Let R denote the model-generated response and G the expert-verified ground truth.

1. Disease Identification Score (S_{dis}). Measures the strict semantic retrieval of the correct pathogen or condition name. Let \mathcal{E}_{dis} be the set of normalized disease entities extracted from G .

$$S_{dis}(R, G) = \max_{e \in \mathcal{E}_{dis}} \mathbb{I}(e \subseteq \text{normalize}(R)) \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if the specific disease entity is explicitly present in the response, and 0 otherwise.

2. Safety Score (S_{safe}). Quantifies the model’s ability to avoid "False Reassurance" errors (i.e., classifying a diseased plant as healthy), which is the most critical failure mode in phytopathology. For the subset of diseased samples \mathcal{D}_{pos} :

$$S_{safe} = 1 - \frac{\sum_{i \in \mathcal{D}_{pos}} \mathbb{I}(\text{"Healthy"} \in R_i)}{|\mathcal{D}_{pos}|} \quad (3)$$

A score of 1.0 indicates zero false negatives (no diseased plant was misdiagnosed as healthy).

3. Clinical Utility Score (S_{clin}). A composite metric evaluating the holistic value of the diagnosis. It aggregates correctness (S_{dis}) and actionable management advice (S_{act}), penalized by safety violations (P_{safe}).

$$S_{clin} = \alpha \cdot S_{dis} + \beta \cdot S_{act} - \gamma \cdot (1 - S_{safe}) \quad (4)$$

where S_{act} measures the semantic overlap of remediation keywords (e.g., "fungicide", "pruning") with the ground truth, and weights are empirically set to $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$ to prioritize accurate identification.

4. Visual Grounding Quality (S_{vg}). Evaluates the hallucination rate of visual symptoms. Let V_G be the set of expert-verified visual cues (e.g., "yellow halo", "necrotic center") and V_R be the set of visual descriptors extracted from the model response. We define S_{vg} as the recall of validated cues:

$$S_{vg} = \frac{|V_R \cap V_G|}{|V_G|} \quad (5)$$

High S_{vg} indicates the model is attending to correct symptomological features rather than generating generic crop descriptions.

5. Visual Feature Extraction Efficiency (E). Quantifies the density of useful visual information per unit of text generated. It is defined as the ratio of verified visual cues ($|V_R \cap V_G|$) to the total word count (W_R) of the response:

$$E = \frac{|V_R \cap V_G|}{W_R} \times 100 \quad (6)$$

A higher E score indicates that the model is providing concise, grounded evidence rather than verbose or irrelevant filler text.

A.2 Data Preprocessing and Quality Assurance

To ensure the integrity and robustness of the **PlantInquiryVQA** benchmark, we implemented a rigorous, multi-stage preprocessing pipeline. This pipeline was designed to eliminate redundancy, standardize visual inputs, and audit the dataset for potential biases or labeling errors prior to annotation.

A.3 Data Cleaning and Standardization

Duplicate Detection and Removal: We employed a two-tiered approach to identify and remove duplicate entries. First, exact duplicates were identified using MD5 file hashing. Second, to capture "near-duplicates" (e.g., images that were slightly compressed or resized but visually identical), we utilized Perceptual Hashing (pHash). We calculated the Hamming distance between image hashes and set a threshold of 10 bits to flag near-duplicates.

Standardization: All images were standardized to a resolution of 1024 pixels. To preserve

the biological integrity of the leaf structures, we avoided simple stretching or cropping. Instead, we utilized a **padding method**: images were resized to fit within the target dimensions while maintaining their original aspect ratio, with the remaining area padded with black pixels. This ensures that critical visual features, such as lesion shape and leaf margins, remained undistorted. Finally, files were renamed using a consistent ‘class_index’ schema to facilitate easier handling.

A.3.1 Technical Audits and Bias Analysis

Following standardization, we conducted a series of automated audits to assess image quality and potential dataset artifacts.

Visual Quality Metrics: We computed technical quality indicators for every image to flag low-quality samples.

Blur Detection: We calculated the variance of the Laplacian operator to detect excessive blurriness.

Exposure Analysis: We measured pixel intensity histograms to detect over-exposure (clipping at 255) and under-exposure (clipping at 0).

Color Saturation: We analyzed the mean saturation in HSV space to identify washed-out images. Flagged images were reviewed manually to determine if they retained sufficient diagnostic value.

Latent Space Consistency Check: To identify potential mislabeled samples, we projected all images into a latent embedding space using a pre-trained ResNet-18 model. We applied a K-Nearest Neighbors (KNN, 3) algorithm using cosine similarity. Images where the majority of neighbors belonged to a different class than the query image were flagged as "suspicious" (e.g., a Bacterial Spot image surrounded by Leaf Miner images in the embedding space). This audit identified 12 potentially mislabeled images for expert review.

Background Bias Assessment: To ensure the model learns from plant features rather than background artifacts, we performed a background bias check. We generated binary leaf masks using HSV color thresholding to isolate leaf pixels from the background. We then computed 30-bin hue histograms for the background pixels solely. By comparing the mean background hue distributions across classes, we confirmed that no specific disease class was strongly correlated with a unique background color (e.g., blue tarps vs. brown soil), which mitigates the risk of "Clever Hans" effects where models cheat by relying on background

Metric	Value
Total QA Pairs	138,068
Unique Images	24,964
Avg. Questions per Image	5.53
Crop Species	34
Disease Types	116
Crop-Disease Combinations	204
Question Categories	2,350
<i>Class Balance</i>	
Healthy Samples	14,858 (10.76%)
Diseased Samples	123,210 (89.24%)

Table 4: **Dataset Overview and Scale.** The dataset covers 204 real-world crop-disease associations with high-density conversational annotations.

Category	QA Samples	% Total
Disease	120,162	87.0%
Healthy	14,858	10.8%
Senescence	2,832	2.1%
Insect Damage	216	0.2%

Table 5: **Distribution of Biological Conditions.** The taxonomy distinguishes between biotic diseases and non-pathological states like senescence and pest damage.

cues. Through this systematic filtering process, we eliminate all corrupted samples, resulting in a curated pool of 152,783 images spanning 34 crop species and 116 disease categories. From this collection, we select 24,964 images to construct the final PlantInquiryVQA dataset.

A.4 Dataset Analysis

We present a comprehensive statistical analysis of the PlantInquiryVQA benchmark, highlighting its scale, biological diversity, and conversational depth. The dataset comprises **138,068 QA pairs** grounded in **24,964 unique images**, representing the largest publicly available CoI dataset for plant pathology (Table 4 and Table 5). The dataset spans **34 crop species** and **116 disease types**, resulting in **204 valid crop-disease combinations**. Further analysis of the dataset is provided below.

Granularity of Disease Severity. Unlike standard classification benchmarks that treat disease as a binary state, **PlantInquiryVQA** provides fine-grained severity annotations for all diseased samples. As shown in Table 6, the data follow a natural distribution: **Mild** cases (20.3%) represent early detection scenarios; **Moderate** cases (55.6%) reflect the most common field presentation; and **Severe** cases (24.1%) represent critical intervention

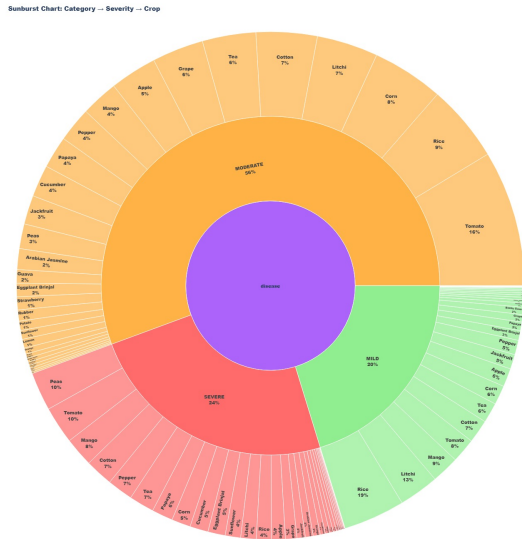


Figure 6: Distribution of diseased samples in the PlantInquiryVQA benchmark, shown as a hierarchical breakdown from disease category to severity level and crop species. The dataset is dominated by Moderate cases, with substantial representation of Severe and Mild conditions across diverse crops, mirroring real-world plant pathology distributions. This structure motivates severity aware Chain of Inquiry trajectories, where diagnostic focus shifts with visual disease progression.

Severity Level	Samples	% of Diseased
Mild	24,332	20.3%
Moderate	66,838	55.6%
Severe	28,992	24.1%
<i>Total Annotated</i>	<i>120,162</i>	<i>100.0%</i>

Table 6: **Severity-Level Annotations.** Granular severity labels enable the model to perform prognostic modeling beyond binary detection.

1027 stages. This granularity enables models to perform
 1028 prognostic reasoning, predicting disease evolution
 1029 based on visual severity cues.

1030 **Conversational Complexity.** To support robust
 1031 multi-hop reasoning, the dataset emphasizes deep
 1032 conversational trajectories. Table 7 illustrates that
 1033 **77.9%** of images are associated with chains of 6
 1034 or more questions, with a global average of **5.53**
 1035 **questions per image**. This depth allows for complex
 1036 logical flows, requiring models to track context
 1037 across temporal (progression), spatial (localization),
 1038 and causal (etiology) dimensions. Furthermore, 86%
 1039 of these samples feature explicit visual grounding,
 1040 linking textual tokens to 74,000 unique
 1041 visual attention patterns.

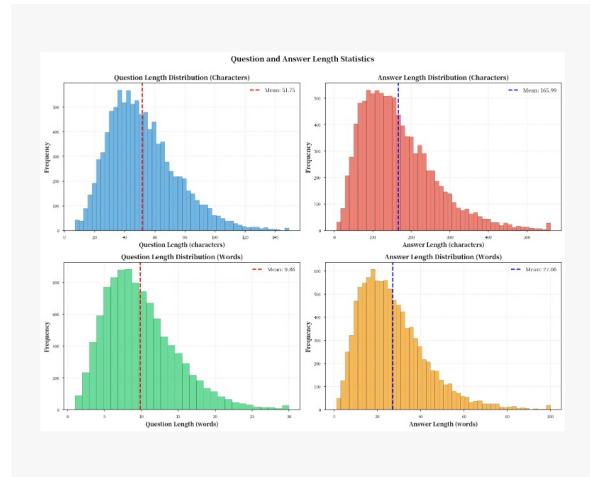


Figure 7: Distribution of question and answer lengths in PlantInquiryVQA, shown in both characters and words. Answers are substantially longer than questions, reflecting the dataset’s emphasis on detailed, multi-step diagnostic reasoning.

Questions per Image	Image Count	% of Dataset
1	2	<0.1%
2	1,277	5.1%
3	1,598	6.4%
4	2,640	10.6%
6	15,971	64.0%
7	3,476	13.9%
Avg: 5.53	Total: 24,964	100.0%

Table 7: **Conversational Depth.** Over 77% of images contain 6 or more questions, facilitating deep multi-turn reasoning and context tracking.

A.5 Visual Cue Extraction Supplementary 1042

A.5.1 Expert Visual Parsing Schema 1043

1044 To ensure high-fidelity visual grounding, we estab-
 1045 lished a standardized diagnostic protocol derived
 1046 from established phytopathology literature [Agrios](#)
 1047 [\(2005\)](#); [Streets \(1972\)](#). Annotators were instructed
 1048 to parse visual evidence across three distinct mor-
 1049 phological dimensions:

1050 **1. Symptomatology and Morphological Character-**
 1051 **ization.** Annotators characterized fine-grained
 1052 attributes of individual lesions to differentiate
 1053 pathogens. Key discriminators included **Lesion**
 1054 **Geometry** (e.g., circular fungal spots vs. vein-
 1055 constrained angular bacterial lesions), **Margin Def-**
 1056 **inition** (e.g., chlorotic halos indicative of toxin
 1057 production or water-soaked bacterial borders), and
 1058 **Textural Features** (e.g., raised galls, powdery
 1059 mycelial growth, or necrotic shot-holes).

2. Spatial Distribution Patterns. Global symptom arrangement provided critical etiological context. The schema required analysis of **Anatomical Preference** (e.g., interveinal, vein-banding, or marginal symptoms) and **Colony Density**, specifically distinguishing between isolated discrete lesions and coalescing necrotic patches that indicate rapid disease progression.

3. Disease Severity Quantification (SAD Methodology). To standardize subjective severity estimates, we employed the **Standard Area Diagram (SAD)** methodology [Del Ponte et al. \(2017\)](#); [Madden et al. \(2007\)](#). Annotators visually compared the total necrotic or chlorotic surface area of the sample against crop-specific SAD reference templates to estimate the percentage of infected leaf area (S), classifying samples into three intervention tiers:

- **Mild** ($S < 15\%$): Early-stage infection typically requiring monitoring.
- **Moderate** ($15\% < S < 30\%$): Established infection necessitating curative intervention.
- **Severe** ($S > 30\%$): Advanced tissue collapse often triggering salvage or removal protocols.

A.5.2 Model Selection

To empirically select the optimal model for the cue extraction pipeline, we evaluated three candidate VLMs on a stratified sample of 500 images. We developed a composite scoring framework to quantify the quality of generated descriptions across three dimensions: grounding, specificity, and structural completeness.

1. Visual Grounding Score (S_{vg}). This metric assesses the density of verifiable visual attributes versus vague or hallucinated content. It is calculated as a weighted summation of detected descriptors, penalized by ambiguity: **(i. Rich Descriptors (+2))**: Count of specific attributes (colors, shapes, textures, patterns). **(ii. Color Diversity (+3))**: Reward for identifying multi-chromatic symptoms (e.g., "yellow halo around brown spot"). **(iii. Grounding Indicators (+1))**: Explicit references to visual evidence (e.g., "visible," "observed," "located"). **(iv. Penalties)**: Vague terms (e.g., "maybe," "some") incur a -0.5 penalty. Unsupported metric measurements (e.g., "5mm wide") incur a severe -2 penalty to discourage hallucinated precision.

2. Specificity Score (S_{sp}). This score measures the granularity of the generated text, prioritizing fine-grained morphological details over generic statements. Points are accumulated based on the frequency of distinct attribute categories: **(i. Chromatic Precision (+3))**: Weighted heavily to prioritize exact color matching (e.g., "necrotic black" vs. "dark"). **(ii. Morphometric Detail (+3))**: Mentions of relative size or scale. **(iii. Textural Characterization (+2))**: Explicit references to surface topology (e.g., "raised," "powdery," "sunken").

A.5.3 Extracted Visual Cues Examples

Comparative (side-by-side) analysis of the expert identified and Qwen-3VL-4B extracted visual cues for a sample image is shown in [Figure 8](#) and [Figure 9](#)

A.6 Generation logic

Our developed generation pipeline is governed by the following logic.

1. **Condition (c) & Severity (s) Initialization:** The pipeline first identifies the biological state $c \in \{Healthy, Diseased, Senescent, Desiccated\}$. If $c = Diseased$, the severity s modulates the **information density** of the response.

2. **Intent-Driven Module Injection (k):** Unlike static VQA, the dialogue trajectory is dynamically assembled based on the epistemic goal k :

- **Diagnosis (k_D):** For **Mild** cases, injects `differential_verification` and `cross_crop_comparison` modules to focus on early symptom detection and rule out lookalikes. For **Moderate/Severe** cases, it triggers `cause_determination` to identify environmental and pest conditions contributing to the disease spread.
- **Prognosis (k_P):** Activates `temporal_evolution` modules. The model reasons about the disease's past (etiology/environment) and future (spread rate), scaling complexity with s .
- **Management (k_M):** Injects `remediation_strategy` modules. If **Mild**, focuses on monitoring and cultural controls (prevention). If **Severe**, focuses on chemical intervention and "rescue" scenarios (crisis management).

1155 3. Counterfactual & Reasoning Augmentation:
1156 To further enhance complexity, we inject **counterfactual**
1157 turns (e.g., "How would the diagnosis
1158 change if the lesions were water-soaked?") into a
1159 subset the chains, specifically targeting the logic
1160 defined in "Instance Variety" heuristic.

1161 **A.7 Diverse CoI Scenarios**

1162 The **CoI** trajectories accros all 12 distinct scenar-
1163 ios are shown in Figure 10, Figure 11, Figure 12,
1164 Figure 13 and Figure 14.

1165 **A.8 Semantic Accuracy Evolution per Inquiry** 1166 **Step**

1167 Figure 15 presents a comprehensive layer-wise
1168 analysis of semantic accuracy evolution across the
1169 seven-step diagnostic trajectory for all 12 evalu-
1170 ated models. By stratifying performance across
1171 three disease severity levels—*Mild* (Green), *Mod-*
1172 *erate* (Orange), and *Severe* (Red)—we observe a
1173 distinct divergence in reasoning stability. High-
1174 performing models generally exhibit a monotonic
1175 increase in semantic alignment as the inquiry pro-
1176 gresses, confirming the CoI framework’s ability to
1177 refine diagnostic grounding through iterative evi-
1178 dence accumulation. However, this positive trajec-
1179 tory is notably dampened in severe cases, where
1180 extensive tissue necrosis introduces visual ambigu-
1181 ity that limits the efficacy of the refinement process
1182 compared to the linear gains observed in mild in-
1183 fection phenotypes.



EXPERT ANNOTATED VISUAL CUES:

Symptomatology: The leaf lamina is intact with normal morphology and no signs of rugosity or hyponasty. Symptoms manifest as scattered, irregular to angular lesions (approx. 4–7 mm) distributed primarily across the midrib and basal regions. Lesions exhibit brownish-yellow necrotic centers with distinct, slightly chlorotic margins; notably, no wide halos or water-soaked borders are present.

Pattern & Distribution: Spots are discrete but show incipient coalescence in high-density areas, accompanied by localized interveinal chlorosis. The overall background coloration displays light green mottling. No evidence of fungal signs (mycelium/sporulation) or mechanical damage is observed.

Severity Assessment: Infection is established with noticeable photosynthetic area loss, but without leaf collapse or extensive necrosis. By visible assessment approximately 20-40% of the leaf is effected, indicating a moderate (Del Ponte et al., 2017) disease.

EXTRACTED VISUAL CUES FROM QWEN3V

Overall color: Light green with mottling of yellowish-green and slightly darker green patches; some areas appear pale. **Surface texture:** Smooth with fine, visible venation; no roughness, wrinkling, or distortion. **Leaf shape integrity:** Normal, undistorted, with intact margins and lobes; no curling, wilting, or deformation.

LESION/SPOT CHARACTERISTICS

Number: Moderate (approximately 10–20 discrete spots visible).

Size: Small to medium (approximately 2–5 mm in diameter).

Shape: Irregular, somewhat angular to oval, not perfectly circular.

Color: Yellowish-green with some brownish-yellow centers; not uniformly colored.

Pattern: Scattered across the leaf blade, not clustered; appears random but follows vein pathways.

DISTRIBUTION

Primarily distributed across the midleaf region, with some extending toward the base; not concentrated at edges or along veins. Coalescing into small patches, but not yet forming large contiguous areas; primarily localized.

ADDITIONAL VISUAL CUES

Mottled yellowish-green patches of discoloration with darker green intermixed; no uniform yellowing. No dead brown or black areas visible. No water-soaked appearance, no glistening or wet look. No powdery/fuzzy growth observed. No visible holes or tissue loss. Yellowing between veins (interveinal chlorosis): Present in some areas, particularly where lesions are dense; not extensive.

SEVERITY ASSESSMENT

MODERATE: Visible lesions covering approximately 15–30% of the leaf surface, with coalescing patches and noticeable discoloration, but no major necrosis or leaf collapse.

Figure 8: Comparison of extracted visual cues for Litchi



EXPERT ANNOTATED VISUAL CUES:

Symptomatology: The leaf retains structural integrity with no rugosity. Diagnostic symptoms appear as distinct, angular lesions strictly delimited by leaf veins (vein-limited). Lesions manifest primarily as pale yellow (chlorotic) patches on the adaxial surface, with older centers turning necrotic brown. Margins are sharp and defined by vascular boundaries rather than diffuse halos.

Pattern & Distribution: Distribution is strictly interveinal, creating a tessellated or mosaic pattern across the lamina. Lesions are scattered but show incipient coalescence into larger necrotic blocks in high-density regions near the midrib. No external fungal mycelium is visible on the upper surface.

Severity Assessment: Infection is active with significant chlorotic coverage, visually estimated at 20–30% of the leaf area. Functional tissue loss is evident, but vascular collapse is absent, classifying the severity as Moderate (Del Ponte et al., 2017).

Overall color: Mix of dark and yellowish-green; abrupt, patchy yellowing concentrated at margins and veins. **Surface texture:** Smooth with clearly defined veins; no visible wrinkling or distortion. **Leaf shape integrity:** Generally intact structure; minor tissue breakdown at edges/lobes; no severe curling or wilting.

LESION/SPOT CHARACTERISTICS

Number: Moderate (numerous distinct areas visible).

Size: Variable from small to moderate; some broad patches rather than discrete spots.

Shape: Irregular, blotchy, and uneven; tends to follow vein structure.

Color: Predominantly yellow; slight browning at tips indicates early necrosis.

Margins: Clear boundaries; no fuzzy or diffuse halos observed.

DISTRIBUTION

Primarily localized along veins and leaf margins; some central discoloration. **Coverage:** Coalescing patches where smaller yellow areas merge into contiguous zones, particularly at outer edges.

ADDITIONAL VISUAL CUES

Necrotic tissue: Minimal; restricted to slight browning at tips/edges. **Water-soaked appearance:** Absent. **Powdery/fuzzy growth:** Absent. **Holes or tissue damage:** Present (small holes likely due to insects/mechanical damage). **Yellowing between veins (interveinal chlorosis):** Observed; concentrated along vascular pathways and margins.

SEVERITY ASSESSMENT

MODERATE: Significant yellowing and coalescing patches along margins and veins covering distinct areas, but lacking extensive necrosis or total tissue collapse.

Figure 9: comparison of extracted visual cues for Bitter Gourd



(a) Corn: Mild Infection

[SEVERITY STATUS: MILD]

Q1: Can you describe the disease we're seeing here?

A: We see distinct, discontinuous yellow streaks running parallel to the veins. This is characteristic of early Maize Streak Virus (MSV).

[Cat: **Diagnosis** | Ground: **Yellow streaks, parallel orientation**]

Q2: How does this look different from Gray Leaf Spot?

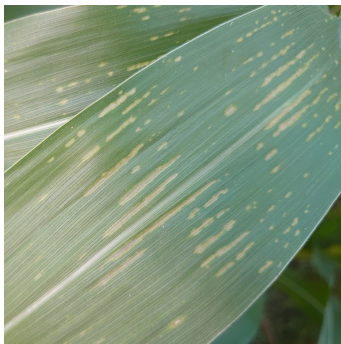
A: Gray Leaf Spot forms rectangular, gray-to-tan lesions. These are continuous yellow streaks, not necrotic rectangles.

[Cat: **Differential Diagnosis** | Ground: **Lesion shape, color**]

Q3: How bad is the infection according to your assessment?

A: It is mild. The streaks are sparse and the majority of the leaf surface remains green and photosynthetic.

[Cat: **Severity Assessment** | Ground: **Lesion density**]



(b) Corn: Moderate Infection

[SEVERITY STATUS: MODERATE]

Q1: What's causing these yellow stripes on my corn leaves?

A: The Maize Streak Virus has progressed. The insect vector (leafhopper) has likely spread the virus, causing these chlorotic bands.

[Cat: **Causal Reasoning** | Ground: **Chlorotic bands**]

Q2: How severe does this look? Is it worth treating?

A: This is moderate severity. The streaks are coalescing. You must control the leafhopper vector immediately to save the yield.

[Cat: **Actionable Severity** | Ground: **Coalescence**]

Q3: Will these yellow marks get worse if left untreated?

A: Yes. The chlorosis will expand, reducing the plant's ability to photosynthesize, which will stunt the cob development.

[Cat: **Prognosis** | Ground: **Chlorosis extent**]



(c) Corn: Severe Infection

[SEVERITY STATUS: SEVERE]

Q1: How bad does it look? Is it really severe?

A: It is severe. The yellow streaking covers almost the entire leaf surface, leaving very little green tissue for photosynthesis.

[Cat: **Severity Assessment** | Ground: **High chlorosis coverage**]

Q2: Do we need to act fast or can we wait?

A: Urgent action is required, though at this stage, the damage to the current crop's yield is likely irreversible. Focus on preventing spread to neighbors.

[Cat: **Urgent Management** | Ground: **Leaf integrity**]

Q3: Without treating right away, what will happen in the next day or two?

A: The leaves will likely turn necrotic (brown/dead) and dry out completely, leading to plant death or total yield failure.

[Cat: **Consequence Prediction** | Ground: **Necrosis risk**]

Figure 10: **Diagnostic Reasoning across Disease Severity Stages.** This figure demonstrates how **PlantInquiryVQA** adapts its questioning strategy as the infection progresses in Maize Streak Virus. **(a) Mild:** The focus is on *Differential Diagnosis* to distinguish the initial streaks from fungal mimics. **(b) Moderate:** The inquiry shifts to *Vector Control* and *Prognosis* as the infection becomes established. **(c) Severe:** The reasoning transitions to *Damage Assessment* and *Salvage*, acknowledging the critical loss of photosynthetic capability.



(a) Jackfruit: Anthracnose

Q1: How many spots do you see on this jackfruit leaf?

A: Around 10 to 15 spots (moderate). Mostly scattered across midleaf region.

[Cat: Severity | Ground: Number of spots, Distribution]

Q2: What does the central part of these spots look like?

A: Pale pinkish-white to light tan. Characteristic of anthracnose in jackfruit.

[Cat: Lesion Characteristics | Ground: Color of central area]

Q3: Are there any signs of the disease getting worse?

A: No, lesions are well-defined with clear red-brown/purple margins. Not merging yet.

[Cat: Progression | Ground: Marginal color, Lesion coalescence]

Q4: What should we do to manage this anthracnose outbreak?

A: Start by pruning affected leaves/branches. Apply fungicides and ensure air circulation.

[Cat: Management | Ground: Treatment]

Q5: Could environmental factors contribute to this infection?

A: Yes, high humidity and poor air circulation promote spread.

[Cat: Causes | Ground: Environmental conditions]

Q6: What changes do you expect to see in the next couple of weeks?

A: If untreated, spots might grow slightly larger but remain scattered.

[Cat: Future Outlook | Ground: Expected lesion progression]



(b) Grape: Anthracnose

Q1: How many spots do you see on this grape leaf?

A: Around 15 to 20 small/medium spots, mostly tan to brown.

[Cat: Severity | Ground: Number, Size, Color]

Q2: What kind of spots does anthracnose usually make on grape leaves?

A: Small, circular, reddish-brown to black spots. Scattered or clustered near midrib.

[Cat: ID | Ground: lesion characteristics, distribution]

Q3: Is this infection severe for grapes?

A: Moderate infection (15-20 spots). Not severe yet, but needs attention.

[Cat: Severity | Ground: Severity]

Q4: Should I start treating these spots right away?

A: Yes, treat immediately to prevent spread. Use recommended fungicides.

[Cat: Immediate Management | Ground: Treatment, Management]

Q5: What caused these spots on my grape leaves?

A: Likely Colletotrichum spp. fungi. High humidity/poor air favor growth.

[Cat: Causes | Ground: Causal agent, Environmental factors]

Q6: What changes might we expect in the next couple of weeks?

A: Spots may grow and potentially merge, leading to extensive damage if untreated.

[Cat: Future Outlook | Ground: Expected progression, Monitoring]

Figure 11: **Cross-species Occurrence of Anthracnose.** The figure illustrates how **PlantInquiryVQA** adapts its CoI to host-specific manifestations of the same pathogen (*Colletotrichum* spp.). **(a) Jackfruit:** The dialogue identifies the classic "bird's-eye" lesions (pale centers, dark margins) and recommends mechanical intervention (pruning) suitable for tree canopies. **(b) Grape:** The dialogue identifies smaller, necrotic tan/brown spots typical of vine infections and prioritizes immediate chemical suppression to prevent rapid merging. This highlights the model's ability to generate context-aware reasoning that reflects the distinct biological symptoms and management protocols unique to each crop species.



(a) Mango: Gall Midge

Q1: What crop is showing these little brown spots?

A: It's mango trees... spots are small, round, and mostly found on the top side.

[Cat: Crop ID | Ground: size, shape, location]

Q2: Can you describe the spots on the leaves?

A: The spots are small, brown, round... maybe just a bit bigger than a pinhead.

[Cat: Symptoms | Ground: size, color, shape]

Q3: How does this compare to other diseases I might see?

A: Different from powdery mildew (no white powder) and bacterial leaf spot (no water-soaked edge).

[Cat: Differential Diagnosis | Ground: comparison, edge]

Q4: Could you explain what causes these spots?

A: Caused by the gall midge insect laying eggs... larvae form bumps that turn brown.

[Cat: Causes | Ground: insect, egg-laying]

Q5: Are these spots going to get worse?

A: Damage looks minor. Won't likely spread much more; tree should recover.

[Cat: Prognosis | Ground: spread, severity]

Q6: Do you notice anything else unusual besides the spots?

A: Everything else seems normal. No wilting, curled leaves; texture is smooth.

[Cat: Additional Cues | Ground: texture, wilting]

Q7: What can I do if I want to prevent these spots?

A: Ensure air circulation, keep area clean, avoid overhead watering.

[Cat: Prevention | Ground: air_circulation, cleanliness]



(b) Mango: Sooty Mold

Q1: How widespread is this sooty mould on my mango leaves?

A: Moderately spread across many parts... covering a good portion of surfaces.

[Cat: Severity | Ground: overall color, surface texture]

Q3: What are the common symptoms of Sooty Mold in mango?

A: Black, fuzzy coating on top side... rubs off easily... associated with honeydew.

[Cat: ID | Ground: colors, patterns]

Q2: Can you tell me how to control this on my mango trees?

A: Wash leaves with water... apply neem oil/soap (often follows insect infestation).

[Cat: Management | Ground: lesion characteristics, distribution]

Q4: Why did my mango leaves get these spots?

A: Likely insect infestation producing honeydew... colonized by fungi (e.g., aphids/scale).

[Cat: Causes | Ground: lesion characteristics, distribution]

Q5: What changes can I expect in the next few days?

A: Mould will continue to grow if conditions remain favorable. Check regularly.

[Cat: Future Outlook | Ground: distribution, lesion characteristics]

Q6: Is there anything else I need to worry about?

A: Obscures photosynthesis... ensure you manage underlying pests to prevent future issues.

[Cat: Impact/Management | Ground: leaf condition]

Figure 12: **Multi-disease Occurrence within a Single Crop Species.** The figure demonstrates distinct CoI trajectories for different pathologies affecting the same host (Mango). **(a) Gall Midge:** The dialogue focuses on structural damage (raised bumps), ruling out fungal pathogens via differential diagnosis, and identifying the insect vector. **(b) Sooty Mold:** The dialogue identifies a superficial fungal issue ("rubs off"), links it to secondary insect activity (honeydew), and focuses on cleaning the leaf surface to restore photosynthesis. This validates that **PlantInquiryVQA** generates context-specific reasoning chains even when the crop type is identical.



(a) Peach: Bacterial Spot

[EPISTEMIC INTENT: DIAGNOSIS]

Q1: What disease is affecting this peach leaf?

A: This is Bacterial Spot (*Xanthomonas campestris*). It appears as small, angular lesions that eventually cause the tissue to drop out.

[Cat: Identification | Ground: Angular spots, tissue loss]

Q2: How can I tell this isn't fungal "shot hole" disease?

A: Bacterial spots are typically more angular and concentrated near the leaf tip/margins. Fungal shot holes usually have distinct purple halos.

[Cat: Differential Diagnosis | Ground: Lesion shape, margins]

Q3: What specific visual cue confirms bacteria?

A: The "shot-hole" effect where dead tissue falls out, leaving ragged holes, without the presence of fungal spores.

[Cat: Symptomatology | Ground: Holes, ragged edges]



(b) Guava: Nutrient Deficiency

[EPISTEMIC INTENT: PROGNOSIS]

Q1: Do you think this will get worse if left untreated?

A: Given it's mild now, if we address nutrient levels, it should improve. Without intervention, it could spread to younger leaves.

[Cat: Early Stage Prognosis | Ground: Current severity, distribution]

Q2: What changes do you expect in the next few weeks?

A: You might see interveinal chlorosis intensifying. Older leaves may turn completely yellow and drop prematurely.

[Cat: Future Outlook | Ground: Chlorosis progression]

Q3: Are there any signs of the deficiency getting severe?

A: Not yet. The spotting is scattered and the leaf structure is intact. Severe cases would show necrotic browning.

[Cat: Progression | Ground: Leaf integrity, spotting density]



(c) Cauliflower: Downy Mildew

[EPISTEMIC INTENT: MANAGEMENT]

Q1: Is it too late to treat these Downy Mildew spots?

A: It's not too late, but act quickly. Damage is severe/irreversible, but treatment can stop spread to the head.

[Cat: Urgent Treatment | Ground: Severity, leaf condition]

Q2: What should I have done differently to prevent this?

A: Treated early when small pale spots appeared. Improved drainage and air circulation would have prevented this.

[Cat: Prevention | Ground: Lesion characteristics]

Q3: If I don't treat this immediately, what happens?

A: The leaf will turn completely brown and die within days. The infection will likely spread to the edible curd.

[Cat: Consequence Analysis | Ground: Leaf condition]

Figure 13: **Evolution of Epistemic Intent across Disease Severity.** The figure illustrates how the CoI shifts its reasoning goal based on the visual status of the plant. **(a) Diagnosis:** In the early/mild stage (Peach), the focus is on *Identification* and distinguishing symptoms from lookalikes. **(b) Prognosis:** In the mild/chronic stage (Guava), the inquiry shifts to *Predicting* the trajectory of the condition (recovery vs. decline). **(c) Management:** In the severe/terminal stage (Cauliflower), the reasoning pivots entirely to *Intervention*, focusing on salvage, containment, and future prevention strategies.



(a) Cherry: Healthy Control

[CONDITION: HEALTHY]

Q1: What crop is shown in this image?

A: Cherry.

[Cat: Crop ID]

Q2: Is this leaf healthy or diseased?

A: Healthy. It shows uniform green color, intact structure, no lesions or spots.

[Cat: Health Status | Ground: Uniform green color, intact structure]

Q3: How does this differ from one affected by fungal diseases?

A: Fungal diseases cause spots/lesions. The absence of these and presence of uniform green confirm this is healthy.

[Cat: Verification | Ground: Absence of lesions]

Q4: What visual signs would indicate disease on this Cherry leaf?

A: We would look for discoloration or abnormal growth. This leaf shows none of these indicators.

[Cat: Symptom Indicators | Ground: Intact tissue]



(b) Senescence / Dryness

[CONDITION: ABIOTIC STRESS]

Q1: Can you tell me what's causing these leaves to turn brown?

A: Due to senescence or dryness. The whole leaf is brown, dry, and papery.

[Cat: ID | Ground: Uniform brownish-tan, papery texture]

Q2: How severe is this condition right now?

A: Quite severe. Entire leaf is dry with no sign of recovery.

[Cat: Severity | Ground: Entire leaf discolored]

Q3: What might have led to such a severe state?

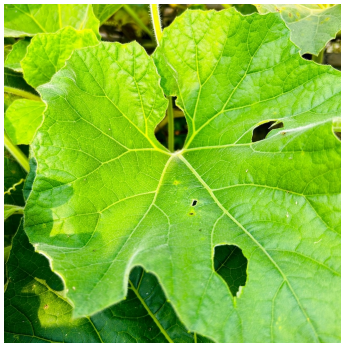
A: Likely lack of water, poor soil, or heat stress. Uniform browning suggests environmental stress.

[Cat: Cause | Ground: Uniform surface affection]

Q4: Do we need to act quickly to stop this from getting worse?

A: Yes. Without treatment, the leaf will drop off soon.

[Cat: Urgency | Ground: Dryness, no partial vitality]



(c) Bottle Gourd: Pest Damage

[CONDITION: PEST INFESTATION]

Q1: How much damage does this look like it's doing?

A: Moderate damage. Leaves are mostly healthy but have holes and curling edges.

[Cat: Severity | Ground: Leaf condition, holes]

Q2: Could these holes be from something other than pests, like sunburn?

A: No, these are clearly pest holes. Sunburn causes scorch marks, not neat-edged openings.

[Cat: Differential Diagnosis | Ground: Holes vs scorch marks]

Q3: What kind of treatment do you suggest?

A: Use vegetable-formulated insecticide or organic neem oil.

[Cat: Management | Ground: Lesion characteristics]

Q4: Do you think the damage will get worse if I don't do anything?

A: Yes, larger lesions may grow and affect more surface area.

[Cat: Prognosis | Ground: Lesion distribution]

Q5: How can I prevent this from happening again?

A: Crop rotation, row covers, and removing fallen debris.

[Cat: Prevention | Ground: Causes]

Figure 14: **Beyond Pathogenic Disease: Healthy, Abiotic, and Pest Conditions.** This figure illustrates the dataset's coverage of diverse plant health states. **(a) Healthy Control:** The model validates health by citing "uniform green color" and the absence of lesions. **(b) Senescence:** The inquiry identifies abiotic stress (aging/dryness) based on global uniform browning and papery texture, distinguishing it from focal infection. **(c) Pest Damage:** The reasoning chain differentiates physical damage (holes) from pathogenic spots and recommends insect-specific treatments (Neem oil) rather than fungicides.

Semantic Accuracy Evolution by Question Number - All Models

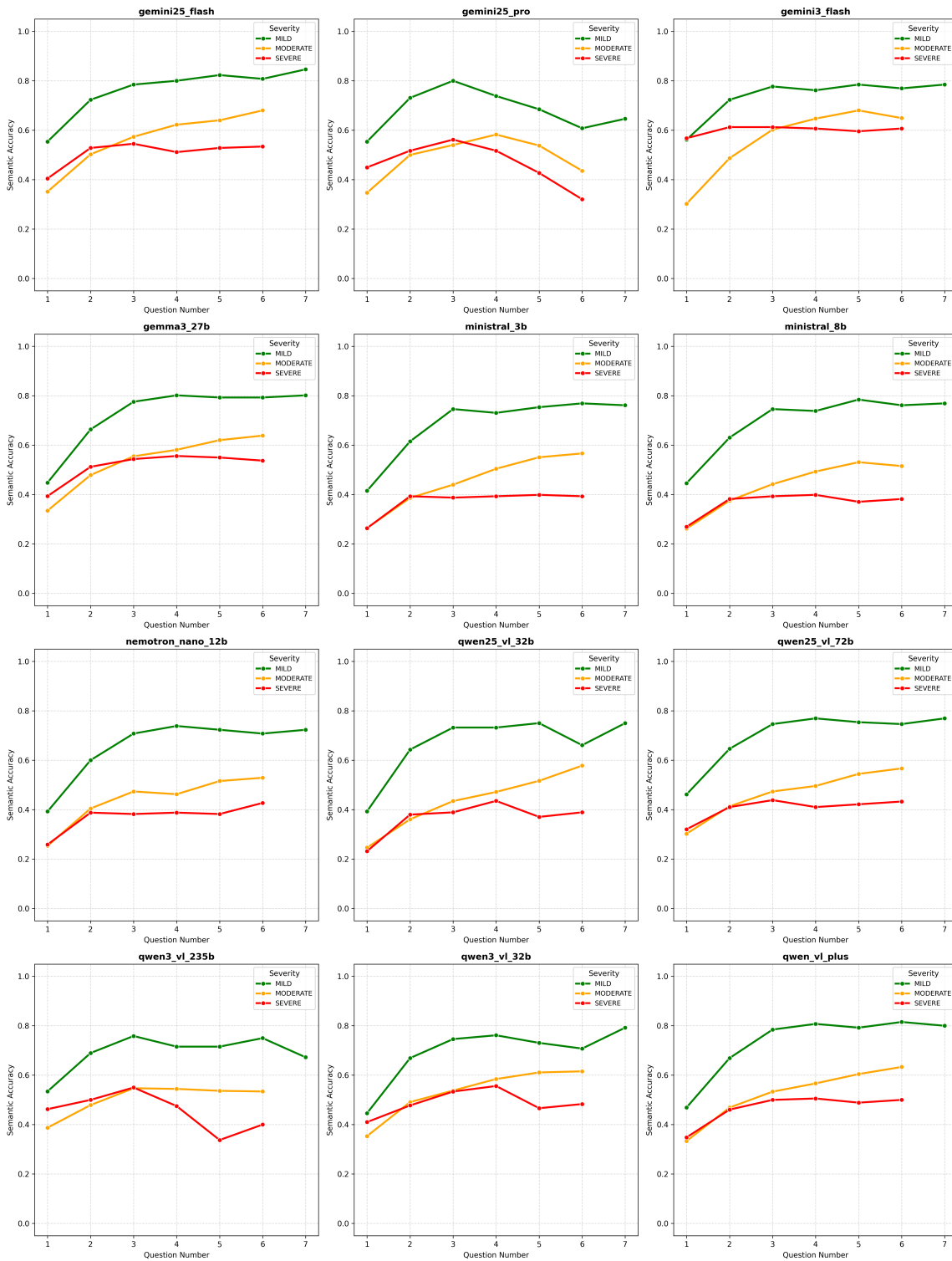


Figure 15: Semantic Accuracy Evolution across the Chain-of-Inquiry Trajectory. The figure illustrates the layer-wise diagnostic accuracy improvement for all 12 evaluated models as they progress through the 7-step diagnostic inquiry. Green lines indicate Mild infection, showing the strongest positive trajectory, while Red lines (Severe) indicate lower baselines and higher volatility. We observe a consistent positive reasoning trajectory, where accuracy improves with each subsequent question, validating the hypothesis that structured inquiry refines diagnostic precision over time. Notably, the performance is stratified by disease severity: models consistently achieve the highest accuracy on Mild cases (green lines), where visual symptoms are distinct, but struggle with Severe cases (red lines), where extensive tissue necrosis often obscures the discriminative features required for accurate grounding.



Figure 16: A comprehensive analysis of diverse disease distribution across crops species of the final PlantInquiryVqa dataset