PaddyVLM: An Expert-tuned Vision-Language Model for Paddy Disease Diagnosis*

Arun Kumar, Sangam Kumar Jena , Pandarasamy Arjunan

Indian Institute of Science (IISc)
Bengaluru, Karnataka, India
{arunkumar12, sangamjena, samy}@iisc.ac.in

Abstract

Paddy is one of the most important crops worldwide, but its cultivation suffers significant yield losses due to diseases and pests that are often difficult to accurately identify and manage. Existing models are largely limited to image classification or detection and lack the contextual agricultural knowledge required for reliable diagnosis and effective mitigation. We present PaddyVLM, a domain-adapted vision-language model for paddy crop diagnosis, capable of identifying diseases and pests, assessing severity, and providing actionable recommendation. Built on LLaVA-v1.5-7B-LoRA, our model is trained using PaddyInstruct, a curated instruction-tuning dataset derived from the Paddy Disease and Paddy Pest datasets using open source Large Multimodal Models (LMMs) and Large Language Models (LLMs). PaddyInstruct combines, LLaVA-13B generated descriptions, Mistral-7B generated simple Q&A and multi-turn Q&A pairs, and expert knowledge refinement. Fine-tuning on this dataset equips PaddyVLM with robust fine-grained recognition and context aware reasoning. Experiments show that PaddyVLM substantially outperforms general-purpose LMMs in both disease and pest diagnosis, demonstrating its potential as a practical expert assistant for farmers and agricultural researchers. All code, datasets, and trained models are available at https://github.com/samy101/paddy-vlm.

Dataset	Num of images	Num of classes
Paddy Doctor Disease	10,407	10
Paddy Doctor Pest	5,673	20

Figure 1: Overview of the Paddy Crop Datasets used for expert-tuning.

Introduction

Agriculture is central to global food security, with paddy (rice) over half of the world's population. Farmers, however, face persistent threats from diseases, pests, nutrient deficiency, and environmental stresses intensified by climate change. Rice health issues alone cause an estimated 10-30% yield loss annually, with severe outbreaks reaching up to 50% (Petchiammal et al. 2023). Early and accurate diagnosis is therefore critical for sustainable and climate resilient paddy farming.

Artificial intelligence (AI) has long supported agricultural tasks such as crop classification, disease detection, and yield estimation. Although deep learning models such as CNNs (Krizhevsky, Sutskever, and Hinton 2012) perform well on specific image based tasks, they remain narrow and dataset dependent, lacking the flexibility required for real world decision support. Moreover, they are typically limited to detection and identification and fail to provide explanations or actionable recommendation needed by farmers and practitioners. Explanatory outputs are particularly critical in agriculture, where farmers and agronomists must understand the cause, confidence, and management reasoning rather than receiving only categorical predictions. Recent advances in vision-language models (VLMs) (Li et al. 2022; Radford et al. 2021) and large multimodal models (LMMs) (Zhang et al. 2024) offer more generalizable capabilities. By combining visual perception with linguistic reasoning, these models can interpret crop symptoms, assess severity, recommend actionable management practices, and answer domain specific queries, making promising tools for agricultural intelligence.

In this work, we introduce PaddyVLM built on LLaVA-v1.5-7B-LoRA, a vision–language model for comprehensive paddy health analysis, covering both disease and pest diagnosis along with farmer oriented recommendations. Inspired by AgroGPT (Awais et al. 2024). However, most agricultural datasets lack textual descriptions, expert dialogue, or contextual reasoning information. PaddyInstruct bridges this gap by synthesizing image-grounded descriptions, structured Q&A, and expert style conversations enriched with reliable agricultural knowledge sources. To effectively adapt the model to paddy cultivation, we construct PaddyInstruct, a comprehensive multimodal instruction-

*Accepted at the First International Workshop on AI in Agriculture (Agri AI), co-located with AAAI 2026.

INSTRUCTION-FOLLOWING EXAMPLE

ATTRIBUTES

TYPE: Leaf
PLANT: Paddy
DISEASED: Yes
NAME: Blast

IMAGE DESCRIPTION

The image depicts a blast disease affecting paddy rice crops. The symptoms of the disease are visible as spindle-shaped leaf lesions on the plants. These lesions are typically indicative of the fungal pathogen *Magnaporthe oryzae*, which thrives in conditions of high humidity and rainfall. To control this disease, resistant varieties of rice and the use of fungicides can be effective measures.

EXTERNAL KNOWLEDGE

Symptoms

- Initially, small, dark-green and water-soaked streaks on interveins from tillering to booting stage
- The progress of the streaks is longitudinal, limited by the veins and soon turn yellow or orange brown.
- All along the streaks bacterial exudates could be observed as tiny yellow or amber colored droplets.

Management Strategies

- The disease can be controlled by proper application of fertilizers and proper plant spacing, use of resistant varieties, and hot water treated seeds.
- Practicing field sanitation is important.
- Ratoons, straws and volunteer seedlings left after harvest can be destroyed to minimize the initial inoculum at the beginning of the season.

Figure 2: An illustration of instruction-following examples used to generate expert-tuning from vision-only agricultural data.

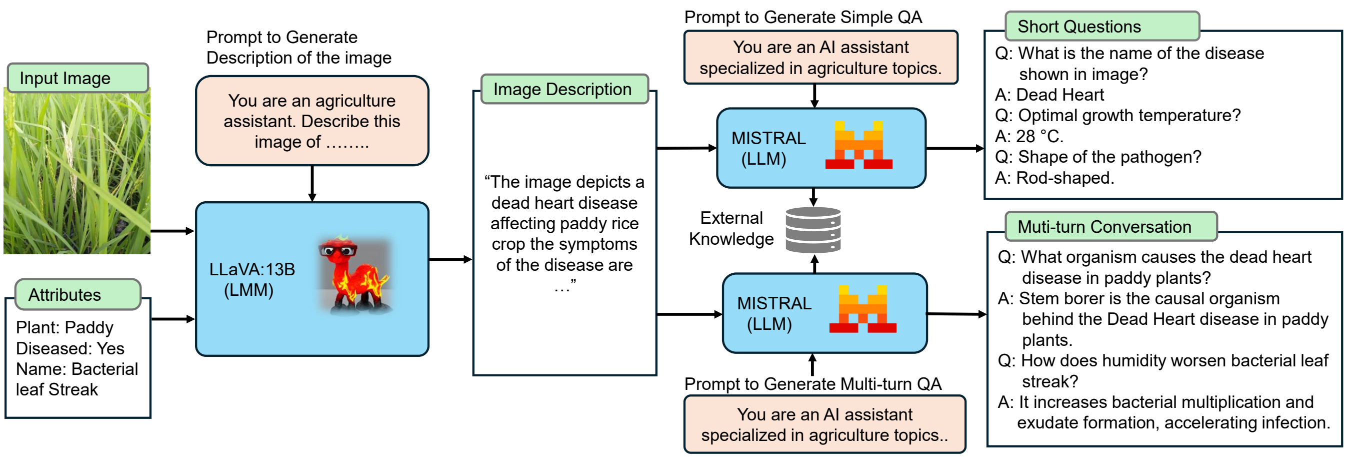

Figure 3: Pipeline used to generate the expert-tuning dataset for PaddyVLM, integrating both disease and pest datasets.

tuning dataset derived from the Paddy Disease and Paddy Pest datasets (Figure 1). The dataset incorporates expert driven annotations to ensure wide coverage of diverse biotic stresses and realistic field conditions. Leveraging this dataset, PaddyVLM learns to align visual features with agricultural domain knowledge through multimodal learning and instruction tuning, enabling it to generate accurate, and context-aware outputs. This positions PaddyVLM as a promising step toward farmer centric, climate resilient decision support systems capable of supporting practical agricultural decision making.

The main contributions of this work are threefold. First, we introduce PaddyInstruct, a large-scale agricultural vision–language dataset consisting of 80k image–text question answer pairs generated from paddy image only datasets using open source large language models (LLMs) and vision language models (VLMs) with carefully designed prompts. The dataset embeds domain knowledge, supports contextual reasoning, and enables robust understanding of paddy diseases and pests. Second, building on this dataset, we develop PaddyVLM, a domain specific vision language model tailored for paddy disease and pest diagnosis. Third, we rigorously evaluated the model against multiple baseline systems, including 13 recent open source VLMs. These contributions collectively advance agricultural multimodal AI from simple recognition toward reliable and field ready decision support.

PaddyVLM: Expert-Tuning and Model Training

PaddyVLM is built through a structured pipeline (Figure 3) aimed at adapting general purpose multimodal models to the specific needs of precision agriculture. Since agricultural datasets are predominantly image only and lack tex-

Table 1: Performance comparison of PaddyVLM with open-source LMMs on the Paddy Disease dataset for Identification and Classification tasks (Accuracy %).

Model	Identification	Classification
LLaVA-7B (Liu et al. 2023a)	57.22	11.56
LLaVA-13B (Liu et al. 2023a)	56.78	10.11
LLaVA-34B (Liu et al. 2023a)	85.56	6.00
LLaVA-Next-8B (AI 2025)	51.00	9.33
Qwen3-VL-8B (Bai et al. 2024b)	42.44	0.00
Qwen3-VL-32B (Bai et al. 2024b)	61.89	0.78
Qwen2.5-VL-7B (Bai et al. 2024a)	15.44	0.33
Qwen2.5-VL-72B (Bai et al. 2024a)	20.67	0.00
Gemma3-4B (DeepMind 2024)	95.78	10.78
Gemma3-12B (DeepMind 2024)	79.67	11.89
Granite3.2-Vision (Research 2025)	17.67	**59.11**
LLaVA-Phi3-3.8B (Liu et al. 2024)	20.22	8.00
MiniCPM-V (Team 2024)	52.56	8.56
PaddyVLM (ours)	**96.5**	**87.0**

Table 2: Performance comparison of PaddyVLM with open-source LMMs on the Paddy Pest dataset, evaluating both Identification and Classification tasks (Accuracy %).

Model	Identification	Classification
LLaVA-7B	27.07	4.46
LLaVA-13B	40.60	5.35
LLaVA-34B	51.13	4.10
LLaVA-Next-8B	10.90	3.57
Qwen3-VL-8B	54.14	12.12
Qwen3-VL-32B	48.50	12.48
Qwen2.5-VL-7B	12.78	8.02
Qwen2.5-VL-72B	28.20	14.44
Gemma3-4B	**88.16**	3.74
Gemma3-12B	70.86	9.98
Granite3.2-Vision	27.26	12.83
LLaVA-Phi3-3.8B	10.34	4.99
MiniCPM-V	29.51	8.73
PaddyVLM (ours)	**95.2**	**81.83**

tual annotations, we use a self-instruct pipeline inspired by AgroGPT (Awais et al. 2024) to generate three types of question-answer pairs: (1) high quality image descriptions,(2) simple Q&A pairs and (3) multi-turn Q&A pairs using image attributes and external agricultural knowledge (Figure 2, 3). We used two complementary datasets for expert-tuning: the Paddy Doctor Dataset (Petchiammal et al. 2023) with 10,407 images across 10 disease categories, and the Paddy Pest Dataset with 5,673 images covering 20 major pest species (Figure 1). The disease dataset provides high resolution images capturing subtle stress patterns, while the pest dataset includes diverse insect species such as Brown Planthopper, Rice Stem Borer Larva, and Sogatella Furcifera, ensuring broad coverage of real world field conditions. These three types of question–answer pairs are designed to capture both fine-grained visual understanding and rich contextual reasoning. Image descriptions are generated using LLaVA-13B (Liu et al. 2023a), providing detailed contextual narratives of each image, including the type of disease or pest, visible symptoms, and characteristic features. Simple Q&A pairs, produced using Mistral-7B (Jiang et al. 2023), focus on direct identification tasks such as determining whether a plant is diseased or identifying the specific disease present. To simulate realistic field interactions, multi-turn Q&A dialogues are also generated using Mistral-7B, capturing expert farmer style conversations involving disease identification, symptom explanation, and management strategies. Prompts are enriched with agricultural knowledge from TNAU (Tamil Nadu Agricultural University 2014) to ensure domain specificity and reliability. Together, these outputs constitute the PaddyInstruct dataset, consisting of detailed image descriptions, structured Q&A pairs, and conversational samples (Figure 3), enabling the model to reason effectively and provide expert level responses.

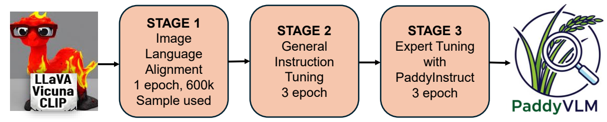

Figure 4: An overview of PaddyVLM three-stage training process .

For the model backbone, we adopt LLaVA-v1.5-7B (Liu et al. 2023b) and fine-tune it using LoRA for efficient adaptation. The combination of CLIP ViT-L/14 (Radford et al. 2021) and Vicuna-7B (Zheng et al. 2023) provides strong multimodal reasoning while remaining computationally manageable. Training follows three stages: (1) vision–language alignment (Liu et al. 2023a), where the model is pretrained on a large collection of image–text pairs, enabling it to learn the correspondence between visual content and linguistic descriptions; (2) generic instruction tuning, where the model is trained on instruction based image–text datasets, allowing it to understand various task instructions such as summarization and question answering; and (3) domain-specific expert tuning, where we fine-tune the model on PaddyInstruct to adapt it to agricultural knowledge and paddy field scenarios (Figure 4). Through these stages, the model progressively improves its ability to jointly interpret images and text, follow complex instructions, and specialize in the target domain. Fine-tuning is performed on an NVIDIA RTX A6000 GPU, with preprocessing carried out on an RTX 5090.

By integrating curated domain data, efficient adaptation, and staged training, PaddyVLM achieves strong performance in disease and pest recognition, symptom interpretation, and generation of actionable recommendations, making it a practical model for real world precision agriculture.

Experiments And Results

The performance evaluation of PaddyVLM demonstrates the strong impact of domain-specific expert-tuning for agricultural vision–language tasks. This section provides two complementary analyses: a quantitative assessment through overall and class wise accuracy comparisons on both the disease and pest datasets, and a qualitative evaluation focusing on interpretability and expert preference. Together, these

Table 3: Classification accuracy (%) comparison of PaddyVLM with Zero-shot Gemini and base model on paddy disease and pest classification datasets.

Model	Disease	Pest
LLaVA-7B (Baseline)	10.00	13.00
Gemini 2.0 Flash Lite	18.40	14.00
Gemini 2.0 Flash	22.00	17.00
Gemini 2.5 Flash Lite	25.60	21.00
Gemini 2.5 Flash	36.63	23.02
Gemini 2.5 Pro	34.14	30.00
Gemini 3 Flash	47.60	39.40
Gemini 3 Pro	42.57	32.40
PaddyVLM	**87.00**	**81.83**

Table 4: Classification performance of PaddyVLM across Paddy disease classes.

Paddy Disease Classes	Precision	Recall	F1-score
Tungro	0.91	0.93	0.92
Hispa	1.00	0.58	0.74
Downy Mildew	0.67	0.82	0.74
Dead Heart	1.00	0.99	1.00
Brown Spot	0.94	0.95	0.94
Blast	0.83	0.90	0.87
Bacterial Panicle Blight	0.92	1.00	0.96
Bacterial Leaf Streak	0.86	0.97	0.91
Bacterial Leaf Blight	0.60	0.96	0.74
Accuracy	**0.87**	**0.87**	**0.87**
Macro Avg.	**0.86**	**0.90**	**0.87**
Weighted Avg.	**0.89**	**0.87**	**0.87**

results show how targeted fine-tuning can effectively adapt general purpose multimodal models into specialized and reliable tools for precision agriculture.

Trained Model Performance Analysis

To comprehensively evaluate the fine-tuned PaddyVLM, we assess its per class and overall performance on both the Paddy Disease and Paddy Pest datasets after removing the 'normal' class. Tables 4, 5 present Precision, Recall, and F1-score for each category, along with macro and weighted averages that account for class imbalance.

Analysis and Key Findings: Across both datasets, PaddyVLM demonstrates strong and reliable performance. **Disease classification:** The model achieves a weighted F1-score of 0.87, with multiple classes such as dead heart, brown spot, and bacterial panicle blight achieving F1 above 0.90. The high macro average Recall of 0.90 indicates consistent capability across all disease categories, including minority classes, without bias toward majority categories.

Pest classification: Despite higher inter-class similarity and a larger number of categories (20), the model maintains a weighted F1-score of 0.806. Several pests including RiceYellowStemBorer, RiceSkipper, and BrownPlanthopper achieve near perfect performance. Lower performing classes

Table 5: Classification performance of PaddyVLM across Paddy pest classes.

Paddy Pest Classes	Precision	Recall	F1-score
Argiope Spider	0.875	0.875	0.875
Brown Marmorated Stink Bug	0.561	0.920	0.697
Brown Planthopper	0.893	1.000	0.943
Damsel Fly	0.933	0.737	0.824
Green Grasshopper	0.843	0.952	0.894
Lady Bug	1.000	0.444	0.615
Leptcorisa Acuta	0.787	0.889	0.835
Long Jawed Orb Weaver	1.000	0.552	0.711
Rice Black Bug	0.853	0.853	0.853
Rice Brown Planthopper	0.400	1.000	0.571
Rice Grasshopper	0.730	0.794	0.761
Rice Green Bug	0.251	0.394	0.288
Rice Green Stink Bug	0.793	0.885	0.836
Rice Shield Bug Trick	1.000	0.111	0.200
Rice Skipper	0.952	0.976	0.964
Rice Stem Borer Larvea	0.387	0.366	0.413
Rice Stink Bug	1.000	0.438	0.609
Rice White Stem Borer	0.563	0.900	0.692
Rice Yellow Stem Borer	0.948	0.982	0.965
Sogatella Furcifera	1.000	0.750	0.857
Accuracy	**0.818**	**0.818**	**0.818**
Macro Avg.	**0.757**	**0.703**	**0.685**
Weighted Avg.	**0.849**	**0.816**	**0.806**

such as RiceGreenBug and RiceStemBorerLarvea reflect limited sample diversity and close morphological resemblance to other species.

Overall: The results confirm that the LoRA fine-tuning effectively adapts the base vision-language model for domain specific agricultural tasks. PaddyVLM achieves robust generalization across 29 total classes and demonstrates strong reliability for real-world deployment in automated crop disease and pest diagnosis.

Comparison with Gemini

We evaluate PaddyVLM against multiple variants of the Gemini multimodal model family on two agricultural perception tasks: paddy disease recognition and paddy pest identification. The evaluation focuses on fine-grained visual classification under realistic field conditions, where models must interpret subtle visual cues related to disease symptoms and pest-induced damage. As shown in Table 3, PaddyVLM consistently and substantially outperforms all evaluated Gemini variants across both tasks. On paddy disease recognition, PaddyVLM achieves an accuracy of 87.00%, compared to 36.63% for Gemini 2.5 Flash and 47.60% for Gemini 3 Flash. Even the higher-capacity Gemini 3 Pro model attains only 42.57% accuracy, leaving a large performance gap in favor of PaddyVLM. A similar pattern is observed for pest identification, where PaddyVLM reaches 81.83% accuracy, more than doubling the performance of Gemini 2.5 Flash (23.02%) and significantly surpassing Gemini 3 Flash (39.40%).

While Gemini Pro variants results are comparable to the Flash models, these gains remain limited and do not compensate for the lack of domain specific grounding. Moreover, we observe a pronounced efficiency trade-off: Gemini Pro models require approximately 4–5× longer inference

Ground Truth

Disease Name : Dead Heart

User	Can you identify the disease shown in the image ?
PaddyVLM	The image shows a case of **Dead Heart** disease in paddy crop.
Gemini	Based on the image, which shows a panicle of rice with many white, sterile, or 'blanked' ,florets (empty kernels), the disease is most likely **Bacterial Panicle Blight (BPB)** or possibly **Rice Blast (Neck Blast)**.
DeepSeek	The image shows a plant with distinctively colored stem, which is white and brown, surrounded by green leaves. The plant appears to be a type of grass or a similar species, and the stem coloration suggests that it may be affected by a disease. The white and brown coloration of the stem is a visual clue that the plant may be suffering from a **fungal disease**.

Figure 5: Qualitative comparison of PaddyVLM with Gemini and DeepSeek for simple question diagnosis tasks.

You are an AI assistant specialized in agricultural topics. You are provided with the text description of an image of a plant, attributes of the plant (such as name, disease), and common information of the plant. Unfortunately, you don't have access to the actual image. You must generate exactly 3 to 5 pairs of question and answer (Q&A). Each question should begin with "Q:" and each answer with "A:". Do not include any narrative text outside the Q&A pairs.
Instructions:
-Focus on visual details that can be seen in the image (e.g., plant type, symptoms, disease, prevention).
-Do not refer to the 'text', 'context', or 'caption' — behave as if you are only seeing the image.
Do not ask speculative or ambiguous questions.
- Avoid referencing numbers, datasets name or image name.
-Maintain consistent formatting as:
Q1:
A1:
Q2:
A2:
(and so on)
Context:
Image Description: {data["**description**"]}
Attributes: {data["**attributes**"]}
External Knowledge: {**external_knowledge**}

Figure 6: Prompt to generate the Multi-turn Question and Answer pairs.

time per image compared to Flash variants, while still under performing on both the disease and pest datasets. This increased latency substantially reduces their practicality for large-scale analysis or real-time decision support in agricultural settings, where timely responses are often critical.

Qualitative analysis (Figure 5, 8) further reveals that Gemini models tend to capture only coarse indicators of plant stress. In many cases, they misattribute pest-induced damage to fungal infections or conflate visually similar disease classes, reflecting a reliance on generic visual associations rather than agronomic reasoning. In contrast, PaddyVLM consistently produces agronomically coherent predictions, correctly distinguishing between closely related conditions such as Brown Spot and Blast, and accurately

You are an agriculture assistant. Describe this image of a {**class_label**} from the {**dataset_name**} dataset. Use the following attributesfor a more detailed and contextual description:\n{**attributes**}

Figure 7: Prompt to generate the Context Description.

identifying pests like Rice Hispa by grounding predictions in localized visual evidence, including damage geometry and texture patterns.

The strong performance of PaddyVLM is primarily driven by its domain-aligned visual-linguistic representation and instruction tuning tailored to agricultural contexts. By incorporating crop pathology descriptors, pest morphology cues, and field-level semantics during training, PaddyVLM learns discriminative features that Gemini models, despite their scale, do not reliably capture. These results underscore the importance of domain-based multimodal modeling for precision agriculture, where accuracy and inference efficiency must be jointly optimized.

Comparative Inference Analysis

We conducted a zero-shot evaluation comparing PaddyVLM with the baseline multimodal models listed in Tables 1 and 2, using the Paddy Disease and Paddy Pest datasets. Each test split contained balanced samples across all classes, and two tasks were evaluated: (1) Identification (healthy vs affected) and (2) Fine-grained Classification (specific disease or pest category). For quantitative evaluation, each model was provided with the image and the full list of possible class names and asked to generate the predicted class. A prediction was counted as correct only if the output exactly matched the ground-truth label, otherwise it was marked incorrect, and accuracy was computed accordingly. The results in Tables 1 and 2 show that although baseline models perform reasonably well on the identification task, they struggle with fine-grained classification across both datasets. Some models achieve isolated gains, but overall accuracy remains low, indicating limited ability to capture the subtle visual and semantic cues present in agricultural imagery.

Ground Truth
Disease Name : Dead Heart
Symptoms : Central shoot of the tiller dries up and turns brown while outer leaves remain green.
Solution : Use resistant varieties, remove and destroy affected tillers, and apply recommended insecticides.

User	See the image and tell me what type of disease, visible symptoms and possible solution?
PaddyVLM	The image depicts a paddy plant affected by the **Dead Heart** disease. **The central shoot of the plant has dried and browned**, indicating severe damage to the heart of the plant. However, the **outer leaves of the plant remain green**, which suggests that disease is localized to the central shoot. The Dead Heart disease is caused by the Rice stem borer, a pest that thrives in warm climates. The presence of this pest in the field can be identified by the symptoms visible on the plant. To control the spread of this disease and prevent further damage to the crop, **resistant varieties of paddy plants can be used**. Additionally, can be applied to kill the Rice stem borer and protect the healthy parts of the plant from further damage. It is important to monitor the field regularly for signs of infestation and take timely action to prevent significant crop loss
Gemini	The symptoms strongly suggest a **Panicle Disease** affecting grain development. The most probable specific diseases are: **Bacterial Panicle Blight (BPB):** Caused by *Burkholderia glumae*. This is a very strong candidate given the white, sterile florets and the erect posture of the affected panicle (it doesn't bend over from the weight of filled grains). **Rice Blast (Neck Blast):** Caused by *Magnaporthe oryzae*……………
DeepSeek	The image shows a plant with a distinctively discolored stem and leaves. The stem is a **pale green color, while the leaves exhibit a mix of green a lighter, almost white or yellowish hue**. The discoloration is most pronounced on the stem, where it transitions from a healthy green to a pale green and then to a whitish color. The visible symptoms suggest that the plant may be affected by a **fungal disease**.

Figure 8: Qualitative comparison of PaddyVLM with Gemini and DeepSeek for multi-turn question diagnosis task.

In contrast, PaddyVLM delivers consistently high performance on both tasks, outperforming all baselines by large margins, exceeding them by more than 40% in fine-grained pest classification. This improvement arises from its domain-specific tuning, which enables reliable differentiation between closely related symptoms such as fungal lesions and pest-induced damage.

Overall, these findings highlight the difficulty general-purpose VLMs face in modeling the nuanced visual semantics of plant pathology and entomology. PaddyVLM's strong zero-shot results demonstrate the effectiveness of domain-grounded multimodal learning for practical crop intelligence and precision agriculture.

Qualitative Results

To complement the quantitative results, we performed a qualitative analysis evaluating the explainability, reasoning quality, contextual awareness, and practical usefulness of the model outputs. This evaluation includes three parts: (1) In-Distribution Qualitative Analysis, (2) Web-based Paddy Disease Analysis, and (3) Agronomist Preference Study. Together, these analyses offer a comprehensive view of the model's real-world performance and interpretability.

In-Distribution Qualitative Analysis: In the simple-question diagnostic task (Figure 5), PaddyVLM correctly identified the Dead Heart and produced a clear, unambiguous response. In contrast, Gemini misclassified the case and produced multiple possibilities, leading to confusion and ambiguity, and DeepSeek provides a generic, narrative-style description that lacks diagnostic grounding.

In the multi-turn setting (Figure 8), PaddyVLM maintained contextual consistency, offering accurate diagnoses and relevant agronomic guidance, correctly identifying symptoms and providing appropriate recommendations. In contrast, Gemini produces multiple uncertain possibilities without clearly identifying symptoms or providing recommendations, resulting in ambiguous outputs. DeepSeek frequently misclassifies both the disease and its associated symptoms and also fails to generate meaningful recommendations.

These findings demonstrate PaddyVLM's superior reasoning, symptom interpretation, and domain-aware dialogue capabilities in in-distribution scenarios.

Qualitative Analysis on Paddy Disease and Pest Images from the Web: To assess real world generalization, we tested PaddyVLM on web sourced paddy disease and pest images containing natural variations such as uneven lighting, cluttered backgrounds, occlusion, and inconsistent image quality. As shown in Figure 9, PaddyVLM demonstrates stronger contextual grounding and fine-grained recognition than general purpose VLMs. For instance, it correctly identified the Hispa pest as "*Hispa (Dicladispa armigera)*," whereas competing models often produced broader or less precise interpretations, their responses were generally less aligned with agronomic context and decision needs.

These results highlight PaddyVLM's ability to distinguish pest-induced and pathological symptoms under noisy field conditions. Its domain aligned reasoning and use of agricultural ontology enable more accurate and reliable diagnostics compared to generic VLMs.

Overall, the combined quantitative and qualitative results confirm that PaddyVLM effectively bridges the gap between

Ground Truth

Disease Name : Hispa.

User	Can you identify the disease shown in the image ?
PaddyVLM	The disease shown in the image is a hispa, known as *Dicladispa armigera*.
Gemini	The image shows rice plants with several small, dark, oblong objects on the leaves, along with some white streaks or marks. Based on these visible symptoms, the most likely issue is: **Pest Infestation, specifically from Planthoppers or Leafhoppers, which can also transmit viral diseases.**
Copilot	From the image, the rice plant looks generally healthy, but I notice a few small dark spots on some of the leaves. These spots could be early signs of a **leaf disease**, though it's not possible to confirm the exact cause just from a photo alone.

Figure 9: comparison of PaddyVLM with Gemini on randomly selected image from web for paddy disease.

multimodal reasoning and domain specific expertise, offering a reliable, and high-performing solution for intelligent paddy disease and pest diagnosis.

Conclusion and Future Work

In this work, we introduced PaddyVLM, a vision–language model tailored for paddy farming that integrates visual recognition with natural language reasoning. Beyond traditional classification, it provides diagnosis, explanation, and actionable recommendations. Our results show that expert-tuning and multimodal learning effectively bridge the gap between advanced AI and field level agricultural needs, supporting more sustainable and climate-resilient farming.

Looking ahead, this approach can be expanded to additional crops and agricultural tasks. Future efforts will focus on scaling expert-tuning to diverse datasets and improving generalizability across farming conditions. Incorporating modalities such as soil data, sensor readings, and climatic indicators will further address challenges posed by climate variability. Ultimately, our goal is to develop a comprehensive decision support framework that enables farmers worldwide to make informed, sustainable, and climate-aware decisions for improved food security.

References

AI, M. 2025. Mapler/LLaMA3-LLaVA-Next: Enhanced Visual Instruction Tuning. Hugging Face Model Card: mapler/llama3-llava-next-8b.

Awais, M.; Alharthi, A. H. S. A.; Kumar, A.; Cholakkal, H.; and Anwer, R. M. 2024. AgroGPT: Efficient Agricultural Vision-Language Model with Expert Tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5687–5696. IEEE.

Bai, J.; et al. 2024a. Qwen2.5-VL: Large Vision-Language Models with Enhanced Visual Grounding. ArXiv preprint arXiv:2409.87654.

Bai, J.; et al. 2024b. Qwen3-VL: Vision-Language Models from Alibaba Cloud. ArXiv preprint arXiv:2407.12345.

DeepMind. 2024. Gemma 3: Multimodal Foundation Models by Google DeepMind. ArXiv preprint arXiv:2406.11220.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.; Lengyel, G.; Lengyel, G.; Saade, A.; Sayed, S.; Balestriero, R.; et al. 2023. Mistral 7B: A Compact and Powerful Language Model. *arXiv preprint arXiv:2310.06825*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*.

Liu, H.; et al. 2024. LLaVA-Phi3: A Compact Multimodal Language Model. ArXiv preprint arXiv:2410.12345.

Liu, W.; Zhang, P.; Li, D.; Li, J.; and Hoi, S. C. 2023b. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*.

Petchiammal, A.; Kiruba, B.; Murugan, D.; and Arjunan, P. 2023. Paddy Doctor: A Visual Image Dataset for Automated Paddy Disease Classification and Benchmarking. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, 203–207.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Research, I. 2025. Granite 3.2 Vision: IBM Foundation Model for Multimodal Reasoning.

Tamil Nadu Agricultural University. 2014. Agritech Portal. https://agritech.tnau.ac.in/. Accessed: 2025-08-22.

Team, O. 2024. MiniCPM-V: Lightweight Multimodal Foundation Model. GitHub repository: https://github.com/OpenBMB/MiniCPM-V.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5625–5644.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Zhang, H.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.