AgriPath: A Systematic Exploration of Architectural Trade-offs for Crop Disease Classification

Anonymous authors
Paper under double-blind review

Abstract

Reliable crop disease detection requires models that perform consistently across diverse acquisition conditions, yet existing evaluations often focus on single architectural families or lab-generated datasets. This work presents a systematic empirical comparison of three model paradigms for fine-grained crop disease classification: Convolutional Neural Networks (CNNs), contrastive Vision–Language Models (VLMs), and generative VLMs. To enable controlled analysis of domain effects, we introduce *AgriPath-LF16*, a benchmark of 111k images spanning 16 crops and 41 diseases with explicit separation between laboratory and field imagery, alongside a balanced 30k subset for standardised training and evaluation. We train and evaluate all models under unified protocols across full, lab-only, and field-only training regimes using macro-F1 and Parse Success Rate (PSR) to account for generative reliability. The results reveal distinct performance profiles: CNNs achieve the highest accuracy on lab imagery but exhibit pronounced degradation under domain shift; contrastive VLMs provide a robust and parameter-efficient alternative with competitive cross-domain performance; generative VLMs demonstrate the strongest resilience to distributional variation, albeit with additional failure modes stemming from free-text generation. These findings highlight that architectural choice should be guided by deployment context rather than aggregate accuracy alone.

1 Introduction

Accurate crop disease detection is critical for mitigating yield loss and supporting food security under increasingly variable environmental conditions. While automated systems can scale diagnostic support, their performance and reliability depend on the context of deployment. Understanding how different model architectures behave in a variety of agricultural contexts remains essential for deploying computer vision systems in this field.

Despite progress, most evaluations focus on narrow architectural families or curated lab imagery. Convolutional Neural Networks (CNNs) (He et al., 2015), contrastive Vision-Language Models (VLMs) (Radford et al., 2021), and generative VLMs (Bai et al., 2025) possess distinct inductive biases and pre-training regimes, yet their comparative performance in complex field environments remains underexplored. Furthermore, existing datasets are frequently crop-specific, domain-imbalanced, and often lack explicit separation between acquisition domains (lab and field). This lack of granularity obscures whether performance reflects genuine visual understanding or domain-specific overfitting, complicating the evaluation of deployment trade-offs.

This work provides a systematic empirical study of model architectures for fine-grained crop disease classification. The primary contributions are:

(1) *AgriPath-LF16:* A domain-aware benchmark featuring 111k images spanning 16 crops and 41 diseases with explicit separation between lab and field acquisition domains. A balanced 30k subset is constructed to support controlled training and evaluation. The dataset is publicly released to facilitate reproducible research[1].

[1]ANONYMOUS LINK

(2) *Multi-Paradigm Comparison:* A unified comparison of CNNs, contrastive VLMs, and generative VLMs is conducted under a shared experimental protocol, incorporating a metric suite of macro-F1 across domains and Parse Success Rate (PSR) to enable direct comparison between deterministic classifiers and free-text generative models.

(3) *Deployment Context Analysis:* Model performance is evaluated under domain-focused training regimes to characterise robustness, in-domain performance, and cross-domain transfer. The analysis highlights distinct performance profiles, clarifying which architectures best suit specific agricultural deployment contexts.

2 Related Work

Advances in computer vision have enabled widespread use of deep learning for image recognition, object detection, and segmentation across various domains. This section surveys prior datasets for crop disease classification, the implications of different computer vision architectures in this task, and outlines the remaining limitations that motivate this study.

2.1 Datasets for Crop Disease Classification

Public datasets for plant disease detection remain limited in scale, environmental diversity, and domain coverage. The widely used PlantVillage dataset introduced by Hughes & Salathe (2016) contains 54,306 images across 14 crops and 38 diseases but is almost entirely composed of lab-captured images with uniform backgrounds, restricting its suitability for evaluating models under real-world conditions. PlantDoc (Singh et al., 2020) provides field-based imagery but at a substantially smaller scale (2,922 samples), limiting its utility for training high-capacity models.

Other datasets such as Plant Pathology 2021 (Thapa et al., 2021) and CornNLB (Wiesner-Hanks & Brahimi, 2019) offer high-quality images but focus narrowly on specific crops and disease categories. Collectively, existing resources exhibit three limitations: (i) strong bias toward lab imagery, (ii) restricted crop and disease diversity, and (iii) insufficient paired lab–field coverage for assessing robustness under distributional shift. To address these gaps, this work introduces AgriPath-LF16, a larger-scale benchmark spanning 16 crops and 41 diseases with explicit separation between lab and field imagery, enabling controlled analysis of performance across heterogeneous deployment contexts.

2.2 CNN-Based Approaches

CNNs have long been the dominant approach for plant disease classification due to their strong inductive biases for local spatial pattern extraction (He et al., 2015). Prior studies demonstrate high performance when training and evaluation occur within the same domain: for example, Islam et al. (2023) report a 98.99% accuracy for ResNet-50 on a PlantVillage subset, and similar lab-based results are found across rice and cotton datasets in K.M. et al. (2023) and Hassan et al. (2022). However, several works highlight severe degradation under distribution shift. Mohanty et al. (2016) show accuracy falling to just above 31% when PlantVillage-trained models were tested on field images.

These findings indicate that CNNs tend to learn representations tightly coupled to their training distribution, resulting in performance variation under heterogeneous acquisition conditions. This trade-off between peak in-domain accuracy and cross-context reliability motivates a controlled comparison against architectures with broader pre-training regimes. Accordingly, this work evaluates CNNs alongside contrastive and generative vision–language models under unified protocols to characterise their relative strengths across diverse deployment contexts.

2.3 Vision–Language Models for Agricultural Tasks

Vision–Language Models (VLMs) combine large-scale visual and textual pre-training, enabling models to align image representations with language-conditioned outputs (Radford et al., 2021). Their application to crop disease detection remains relatively recent. The VLCD framework (Zhou et al., 2024) integrates visual

features with textual disease descriptions, reporting promising results on grape leaf diseases, although its single-crop focus limits broader generalisability.

More recent work, such as AgroGPT (Awais et al., 2024), extends multimodal LLMs toward expert-style reasoning and visual question answering (VQA) in agricultural contexts through the AgroEvals benchmark, emphasising conversational capability and advisory interaction. However, their primary focus does not lie in fine-grained classification across diverse crops and acquisition conditions. Emerging benchmarks such as MIRAGE, introduced by Dongre et al. (2025), evaluate VLMs on information-seeking behaviour and multimodal dialogue with agricultural experts. Similar to AgroGPT, such benchmarks presuppose that models possess reliable visual understanding.

To address these limitations, this work provides a unified empirical comparison of model architectures under various training and evaluation regimes. By examining performance across acquisition contexts, it assesses whether the visual capabilities assumed in advisory and conversational benchmarks translate into reliable fine-grained classification under heterogeneous deployment conditions.

3 Building VLMs for Crop Disease Classification

3.1 Dataset

This study introduces AgriPath-LF16, a large-scale dataset containing 111,307 images across 16 crops, 41 diseases, and 65 crop–disease pairs (Table 3). The dataset spans two acquisition domains: controlled laboratory conditions and natural field environments. Lab images exhibit uniform backgrounds and consistent lighting, whereas field images contain cluttered scenes and variable conditions (Figure 1).

Figure 1: Example of source difference in AgriPath-LF16. **Left**: Lab-sourced image of Tomato with Bacterial Spot. **Right:** Field-sourced image of the same disease, illustrating background clutter and lighting variation.

The full dataset exhibits class imbalance and uneven domain representation, with some crop–disease pairs dominated by a single acquisition source. This imbalance complicates controlled evaluation of robustness under domain shift.

To mitigate imbalance, a balanced subset, AgriPath-LF16-30k, was constructed via class-preserving downsampling. The subset contains 28,482 images, retains all 65 crop–disease pairs, and follows an 80/10/10 train–validation–test split. Balancing prioritised equitable lab–field representation within each class where data permitted. Full downsampling rules are provided in Appendix A.2.

Figure 2 illustrates the distribution of AgriPath-LF16-30k across 65 crop–disease pairs, revealing a substantially flattened class profile. Although minor imbalance persists due to data availability, the subset avoids the severe long-tailed skew typical of agricultural datasets, enabling more reliable cross-class and cross-domain evaluation.

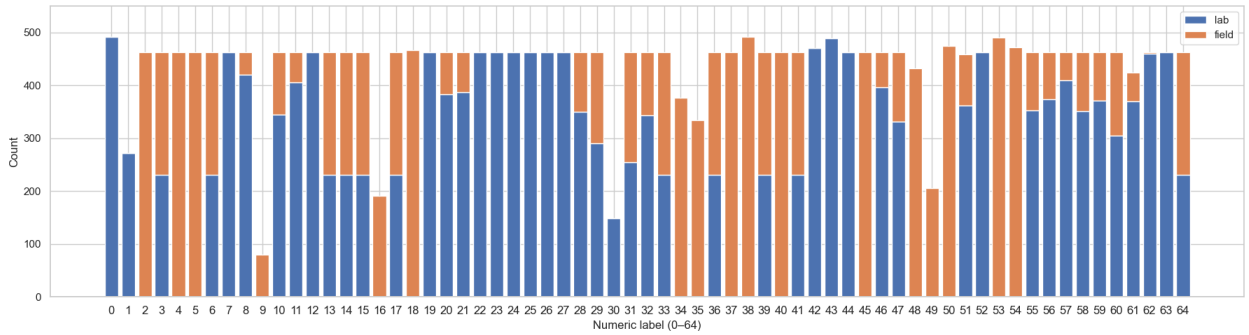


Figure 2: Class and source distribution of AgriPath-LF16-30k across 65 crop-disease pairs. Blue refers to lab-based samples and orange refers to field-based samples

The figure also shows that some classes remain dominated by lab imagery due to limited field data. This reflects practical constraints in real-world data collection rather than a deliberate design choice. Retaining this imbalance allows robustness to acquisition conditions to be evaluated under realistic deployment scenarios.

AgriPath-LF16-30k provides a domain-aware benchmark for evaluating performance, robustness, and generalisation across CNNs, contrastive VLMs, and generative VLMs under both controlled and in-the-wild conditions.

3.2 CNN Baseline

A ResNet-50 model pre-trained on the ImageNet dataset is used to create the CNN baseline. All backbone layers were initialised with ImageNet weights; following previous work on using CNN for classification tasks, early layers were frozen, while only the final residual block group and a newly initialised classification head were fine-tuned (He et al., 2015).

To identify an effective configuration, a small grid of nine experiments was conducted, varying batch size (16, 32, 64) and learning rate (1×10^{-4}, 2×10^{-4}, 5×10^{-4}). Each configuration was trained under three regimes: (a) using the full training set containing both lab and field images, (b) using only lab images, and (c) using only field images. This design provides a consistent set of CNN baselines for assessing in-domain performance as well as robustness under both moderate and extreme forms of distributional shift. Of the 9 experiments, the experiment with the lowest validation loss was selected for analysis.

3.3 Contrastive VLM Experiments

Contrastive VLMs employ Vision Transformer (ViT) backbones pretrained via image–text alignment objectives and perform inference over a fixed label space. Two models were evaluated: Google's SigLIP base model (≈203M parameters) (Zhai et al., 2023) and OpenAI's CLIP ViT-L/14 (≈427M parameters) (Radford et al., 2021).

These contrastive models were evaluated in two settings: 1) *Zero-shot classification* and 2) *linear probing*.

In zero-shot classification, we embed images and textual class descriptions into a shared representation space, with labels assigned via cosine similarity. For each class, descriptive templates were constructed from crop and disease metadata (see Appendix C.1 for the zero-shot templates), and similarity scores were averaged across a small template ensemble.

For linear probing, we train a linear classifier using frozen image embeddings, updating only classifier parameters while keeping the pretrained vision encoder fixed. Linear probing was performed under the same three training regimes as other model types: full dataset, lab-only, and field-only.

All contrastive models were trained with a batch size of 64, with learning rates selected via a small manual sweep including 0.001, 0.003, and 0.01.

3.4 Generative VLM Experiments

In contrast to encoder-only contrastive models, generative VLMs produce free-text outputs conditioned on visual input. Three generative VLMs of varying scale are evaluated: Qwen2.5-VL 3B Instruct, Qwen2.5-VL 7B Instruct (Bai et al., 2025), and SmolVLM 500M Instruct (Marafioti et al., 2025).

Zero-shot performance was evaluated prior to adaptation using three prompting strategies: (1) a simple instruction (Pure), (2) a context-augmented instruction providing the full set of crop–disease pairs (Context), and (3) a multiple-choice format (MCQ) requiring selection from four candidate classes, where three candidates were randomly selected for each sample, and one candidate was the true label.

A frozen-vision (FV) condition was included, where the vision encoder remained fixed and LoRA was applied only to the language components.

To adapt the selected VLMs to the task, Low-Rank Adaptation (LoRA) (Hu et al., 2022) was applied to the models, with all experiments using a multimodal conversational format (Appendix D.2). A Bayesian optimisation sweep tuned learning rate, weight decay, and LoRA rank r (see Appendix D.3 for sweep configurations). A scaling factor of $\alpha = 2r$ was adopted after preliminary experiments showed consistently lower validation loss compared to $\alpha = r$.

Fine-tuning followed the same three training regimes as the CNNs: full dataset, lab-only, and field-only.

4 Evaluation & Analysis

4.1 Evaluation Setup

Three model families are evaluated under a unified protocol: CNNs, contrastive VLMs, and generative VLMs. CNNs and contrastive models produce closed-set predictions, whereas generative VLMs output free-text responses that must be mapped to valid class labels. Performance is evaluated using macro-averaged Precision, Recall, and F1-score across the 65 classes, with F1 reported as the primary metric. Generalisation is assessed by evaluating performance separately on lab, field, and combined test sets, enabling analysis under domain shift. Results for overall performance are reported in Table 1, while generalisation experiments can be found in Table 2.

To ensure fair comparison, generative outputs that cannot be parsed into valid class labels are treated as incorrect predictions via assignment to a dedicated `false_parse` class. Parse Success Rate (PSR) is reported separately to distinguish visual misclassification from instruction-related failures.

4.2 CNN-Based Models under Domain Shift

To calibrate task difficulty, two naïve reference classifiers were evaluated. A Random classifier achieved an F1-score of 0.013, consistent with chance performance, while a Majority-class baseline achieved an F1-score of 0.0006, indicating the absence of a dominant class and confirming the non-triviality of the task.

A ResNet-50 CNN pretrained on ImageNet and fine-tuned for crop disease classification represents a conventional supervised baseline (Islam et al., 2023; K.M. et al., 2023). When trained and evaluated on the combined lab and field dataset, it achieves an F1-score of 0.910, demonstrating that CNN-based models can perform very effectively when training and test conditions are aligned (see Table 4 for hyperparameter sweep). Notably, the same model attains a lab-only F1-score of 0.973, indicating particularly strong performance in controlled imaging environments.

However, this performance is highly sensitive to changes in the data distribution. As shown in Table 2, when evaluated exclusively on field images, the CNN's F1-score decreases from 0.973 (Lab) to 0.703 (Field), corresponding to a relative reduction of $\approx 27.75\%$. Under more extreme mismatch, a CNN trained only on lab images achieves a lab F1-score of 0.968 but degrades to a field F1-score of 0.045, indicating almost complete failure to transfer to in-the-wild conditions ($\approx 95.35\%$ collapse). The converse pattern is also observed: a CNN trained only on field data achieves a relatively strong field F1-score (0.737) but performs poorly on lab images (0.133).

	PSR (%)	F1	Precision	Recall
Random	–	**1.3**	1.4	1.3
Majority	–	0.06	0.03	1.5
ResNet-50	–	**91.0**	92.1	90.7
SigLIP-ZS	–	0.16	0.15	0.64
CLIP/L/14-ZS	–	**14.3**	20.0	19.1
SigLIP	–	90.1	90.4	90.2
CLIP/L/14	–	**91.1**	91.6	91.1
Qwen2.5-VL-3B-ZS-MCQ	21.2	23.7	68.7	16.5
Qwen2.5-VL-7B-ZS-MCQ	94.8	**65.9**	71.0	65.9
SmolVLM-500M-ZS-Context	25.0	0.2	0.4	0.6
Qwen2.5-VL-3B-FV	99.9	87.6	88.2	87.4
Qwen2.5-VL-7B-FV	100.0	**89.8**	90.5	89.9
SmolVLM-500M-FV	99.9	75.3	76.2	75.9
Qwen2.5-VL-3B	99.9	89.0	89.5	88.9
Qwen2.5-VL-7B	99.8	**90.5**	90.9	90.4
SmolVLM-500M	100.0	87.8	89.6	87.5

Table 1: **Overall performance on the combined test set.** Macro-F1, precision, recall, and Parse Success Rate (PSR) are reported in percentage scale (0–100). PSR is applicable only to generative VLMs. Bold values denote the strongest F1-score within each experimental group.

Taken together, these results indicate that CNN-based classifiers learn highly domain-specific visual representations that support excellent performance under controlled conditions but generalise poorly across acquisition settings. This behaviour suggests that CNNs are well suited to laboratory-based disease classification pipelines with standardised imaging protocols, but their reliability degrades substantially when applied to heterogeneous field environments.

4.3 Contrastive Vision-Language Models

Contrastive VLMs are evaluated in both zero-shot and linear-probe settings under the same full, lab-only, and field-only regimes used for CNNs.

Zero-shot contrastive inference performs poorly (F1 $\leq$ 0.15): cosine similarity between image embeddings and templated class prompts does not reliably separate the 65 crop–disease categories. This indicates that web-scale contrastive alignment alone does not encode the fine-grained distinctions required, and that task-specific adaptation is necessary.

Linear probing substantially improves performance, showing that pretrained contrastive encoders already contain disease-relevant features. Under full-dataset training, performance increases with model capacity; relative to CNNs, contrastive models compete on lab images and improve robustness on field images, suggesting that their pre-trained representations are less tightly coupled to controlled imaging conditions.

Under domain-exclusive training, contrastive VLMs degrade on the opposite domain: field-trained models transfer to lab setting better than the inverse. This suggests lab-only training encourages reliance on clean, low-variance cues that do not transfer to real-world settings, whereas field-only training yields classifiers that tolerate cleaner lab imagery. Compared to CNNs, contrastive VLMs exhibit more stable cross-domain behaviour, though substantial degradation persists under extreme lab-to-field mismatch.

Contrastive models operate over a fixed label space with deterministic inference, eliminating parse failures and interface-induced errors. Their closed-set formulation constrains output flexibility but provides predictable and structurally reliable predictions. Overall, contrastive VLMs offer a more domain-tolerant alternative to conventional CNNs while maintaining architectural simplicity and minimal task-specific adaptation.

	PSR (%)	F1(Combined)	F1(Lab)	F1(Field)
Trained on full dataset				
ResNet-50	–	91.0	**97.2**	70.3
SigLIP	–	90.1	93.2	73.7
CLIP/L/14	–	**91.1**	95.6	75.2
Qwen2.5-VL-3B	99.93	89.0	91.8	74.6
Qwen2.5-VL-7B	99.82	90.5	93.0	**78.7**
SmolVLM-500M	100.0	87.8	92.8	70.6
Trained on lab-only dataset				
ResNet-50	–	56.1	**96.8**	4.5
SigLIP	–	56.5	95.0	14.0
CLIP/L/14	–	56.8	95.9	10.8
Qwen2.5-VL-3B	94.44	55.3	88.0	23.4
Qwen2.5-VL-7B	100.0	58.3	94.1	**25.2**
SmolVLM-500M	99.93	**58.9**	94.4	18.3
Trained on field-only dataset				
ResNet-50	–	37.4	13.3	73.7
SigLIP	–	41.3	19.6	75.8
CLIP/L/14	–	41.2	17.4	**77.5**
Qwen2.5-VL-3B	99.16	39.4	15.8	75.0
Qwen2.5-VL-7B	92.05	**43.7**	**20.3**	75.1
SmolVLM-500M	99.86	34.9	10.6	67.5

Table 2: **Generalisation performance across domain-specific training regimes.** Macro-F1 scores (percentage scale) are reported for combined, Lab-only, and Field-only test sets. PSR is shown for generative models. Bold values indicate the strongest F1 within each training regime.

4.4 Generative Vision-Language Models

Generative VLMs extend contrastive alignment by coupling pretrained vision encoders with large language models capable of free-text reasoning and structured output generation. Their behaviour is examined under increasing levels of task adaptation, from zero-shot prompting to supervised LoRA fine-tuning.

In the zero-shot setting, performance is highly sensitive to prompt formulation. Simple instruction prompts yield uniformly low F1-scores ($\leq$0.04), whereas structurally constrained multiple-choice prompts substantially improve performance for larger models (see Table 7 for all zero-shot performance metrics). This contrast indicates that task-relevant visual knowledge is present in pretrained representations, but zero-shot performance is limited by the model's ability to reliably map visual evidence into the required structured output format. Unlike contrastive models, inference quality is therefore jointly determined by representation and interface design.

The frozen-vision (FV) condition isolates the contribution of pretrained visual features by applying LoRA only to language components. Under this constraint, both Qwen variants retain strong classification performance, implying that substantial disease-relevant signal is already encoded in the vision backbone. SmolVLM exhibits greater degradation, suggesting that model capacity influences how effectively frozen visual representations can be exploited.

Under full LoRA fine-tuning on the combined dataset, generative VLMs achieve F1-scores $\geq$0.875, competitive with CNNs. While CNNs achieve higher lab accuracy, VLMs show more balanced degradation across domains. On average, VLMs experience a smaller relative drop under domain shift ($\approx$ 19.35%) compared to the CNN ($\approx$ 27.75%), with robustness improving with model scale.

Cross-domain generalisation is evaluated via domain-exclusive fine-tuning. Lab-only training leads to substantial degradation when evaluated on field data, with varying severity across models. Qwen-based models

exhibit a smaller relative drop ($\approx 74\%$) compared to the CNN's collapse ($\approx 95\%$), retaining a meaningful fraction of their performance under severe mismatch. SmolVLM shows a larger decline ($\approx 80\%$), suggesting that model capacity influences the retention of transferable visual features under extreme domain shift.

As observed with contrastive models, transfer asymmetry emerges, however, inversely; lab-to-field degradation is smaller than field-to-lab. This pattern suggests that web-scale pretraining biases vision encoders toward visually complex, in-the-wild imagery, providing a robustness prior during lab-only adaptation, whereas laboratory imagery constitutes a comparatively narrow distribution.

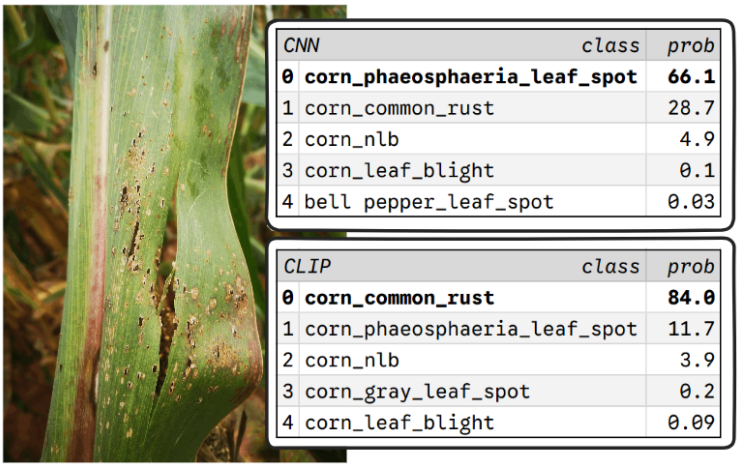
Overall, generative VLMs exhibit strong pretrained visual representations that can be unlocked with modest supervision, though their effectiveness depends on training domain and alignment strategy. Their generative interface introduces additional complexity, but also supports greater robustness under domain shift, reflecting a trade-off between interface simplicity and representational generality.

4.5 Failure Analysis

This section analyses the different failure modes by providing specific examples of these failure modes that are characteristic of each model type.

4.5.1 CNN Failure Modes

To analyse CNN failure modes, the confusion matrix of the ResNet50 model trained on combined lab and field data was examined (Figure 5). The matrix revealed that `corn_common_rust` (Class 13) exhibited the highest error rate, primarily being misclassified as `corn_phaeosphaeria_leaf_spot` (Class 18). Figure 3 illustrates a representative case of this inter-class leakage.



CNN	class	prob
0	**corn_phaeosphaeria_leaf_spot**	**66.1**
1	corn_common_rust	28.7
2	corn_nlb	4.9
3	corn_leaf_blight	0.1
4	bell pepper_leaf_spot	0.03

CLIP	class	prob
0	**corn_common_rust**	**84.0**
1	corn_phaeosphaeria_leaf_spot	11.7
2	corn_nlb	3.9
3	corn_gray_leaf_spot	0.2
4	corn_leaf_blight	0.09

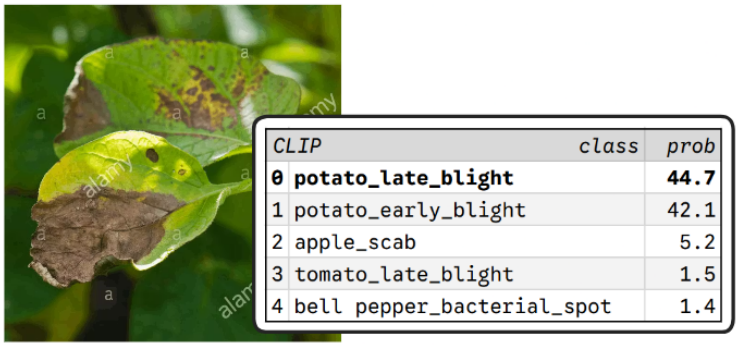
Figure 3: A corn crop with common rust in the field with probabilities for a CNN and CLIP prediction.

The input image features a damaged leaf where necrotic lesions and perforations dominate the local texture, alongside darker rust-coloured pustules distributed across the leaf surface. The CNN's preference for leaf spot ($\approx 66\%$) may stem from an over-reliance on shape-dominant local features under distributional noise, reflecting an inductive bias toward lesion geometry and texture patterns. The VLM, however, aligns the image with higher-level semantic concepts such as rust-like visual cues, resulting in correct identification of the rust disease with high confidence ($\approx 84\%$). Despite this, the non-negligible probability assigned to the correct class ($\approx 28\%$) indicates that the CNN maintains partial recognition of the underlying pathology.

4.5.2 Contrastive VLM Failure Modes

Similar to the CNN, the confusion matrix of the CLIP model was consulted (Figure 6), and an example where the CLIP model exhibits pronounced uncertainty between two closely related disease classes is examined.

Figure 4 shows a potato leaf affected by blight-like symptoms, for which the model assigns comparable probability to `potato_late_blight` (45%) and `potato_early_blight` (42%). Generally, both diseases present overlapping visual phenotypic characteristics, including irregular necrotic regions and diffuse chlorosis, making disambiguation challenging.



CLIP	class	prob
0	**potato_late_blight**	**44.7**
1	potato_early_blight	42.1
2	apple_scab	5.2
3	tomato_late_blight	1.5
4	bell pepper_bacterial_spot	1.4

Figure 4: A potato crop with an ambiguous blight disease in the field with an uncertain CLIP prediction.

Rather than a spurious misclassification, this example reflects difficulty in resolving fine-grained class boundaries under substantial visual overlap. The model's probability distribution reveals semantic proximity between the two classes in the embedding space, indicating that the visual evidence does not strongly favour a single diagnosis. Although this uncertainty is preferable to overconfident error, it remains problematic in high-stakes agricultural contexts, where early and late blight require different interventions. This illustrates a limitation of contrastive VLMs: strong global semantic alignment does not guarantee reliable separation of subtle intra-class distinctions without additional contextual or temporal cues.

4.5.3 Generative VLM Failure Modes

Unconstrained text generation introduces failure modes not present in discriminative models. This analysis examines outputs that could not be parsed into valid crop–disease pairs under the task specification.

Across all generative VLM experiments, 431 outputs failed parsing. Of these, 39.7% were empty generations, representing silent task breakdowns. Among the remaining non-empty failures, 12.69% were classified as hallucinations. Hallucinations occur in several systematic forms. First, models introduce non-existent modifiers or entities (e.g., `icelandic raspberry`, `icicle apple`), producing crop–disease combinations outside the dataset ontology. Second, unsupported crop substitutions appear (e.g., `banana`, `sage`, `hazel`), despite their absence from the training distribution. Third, other failures stem from semantic vagueness rather than outright fabrication. For instance, models can output underspecified classes (e.g., `leaf_spot`, `tomato_spot`) that correspond to multiple dataset classes and cannot be resolved unambiguously.

These findings indicate that although generative VLMs offer expressive and flexible predictions, they introduce non-trivial risks of hallucination and silent failure in high-stakes agricultural deployment.

4.6 Summary of Findings

Results show that crop disease detection depends heavily on the deployment context. CNNs achieve the strongest performance on lab-based data, but experience the steepest degradation under domain shift. Consistent with prior findings that CNN representations are tightly coupled to their training distribution, their reliance on local texture and shape cues makes them sensitive to changes in environmental context, resulting in reduced aggregate performance and systematic class-level confusions under field conditions.

Contrastive VLMs demonstrate improved robustness to domain shift relative to CNNs. Their errors tend to cluster around visually or semantically adjacent classes, indicating uncertainty rather than complete failure. This suggests that contrastive pretraining encourages more transferable visual representations, while still limited by visually subtle, fine-grained disease distinctions.

Generative VLMs are the most robust to distributional shift among the three model types, but introduce an additional failure mode beyond misclassification. Despite their robustness, their open-ended generation objective leads to non-trivial rates of invalid or hallucinated outputs that are critical in safety-sensitive agricultural contexts.

Taken together, these findings indicate that no single model family is universally optimal. Instead, deployment requirements should guide model selection: CNNs are preferred when high lab-based accuracy is the primary objective; contrastive VLMs offer a more robust alternative for fixed sets of crops under moderate domain shift; and generative VLMs are most suitable when flexibility and extended capabilities are required.

5 Conclusions & Outlook

This work presents a systematic comparison of model architectures across realistic agricultural domains and introduces a balanced benchmark for crop disease evaluation. The results show that architectural suitability depends on deployment context rather than peak accuracy.

5.1 Limitations & Future Work

Despite providing a controlled benchmark for cross-domain evaluation, several limitations remain. First, although AgriPath-LF16 reduces extreme imbalance, field-sourced samples remain comparatively underrepresented for certain crop–disease pairs. This limits the ability to fully characterise robustness under real-world variability. Second, the dataset lacks contextual or temporal information that represents disease progression, which is dynamic and ambiguous in isolation. Additionally, multimodal agricultural metadata such as region, climate, soil characteristics, or seasonality were not incorporated. Computational constraints limited exploration of larger architectures, broader hyperparameter sweeps, and extended ablations.

Future work should expand field data collection to reduce domain imbalance and increase environmental coverage. Integrating contextual metadata into generative VLM pipelines may improve reliability and reduce hallucinations through agricultural grounding (Yao et al., 2023). Temporal modelling of disease progression using longitudinal imagery is another promising direction. Moreover, reasoning-aware fine-tuning that promotes localisation of symptomatic regions and suppresses background cues may improve robustness (Bansal et al., 2025). Finally, real-world farmer-facing evaluations would clarify usability, trust calibration, and practical performance beyond benchmarks.

5.2 Statement of Broader Impact

This study supports the development of robust, context-aware disease detection systems to enhance agricultural productivity and food security. Early, accurate identification of plant diseases minimises yield loss, enables timely intervention, and reduces unnecessary pesticide use. Architectures that maintain robustness under domain shift are especially critical for real-world agricultural settings with uncontrolled imaging conditions. Furthermore, generative VLMs offer flexible diagnostic potential by integrating contextual metadata and providing structured reasoning, improving accessibility for smallholder farmers and facilitating integration into precision agriculture workflows.

Generative models may produce hallucinated outputs that could lead to incorrect interventions; mitigation requires structured constraints, grounding, and human oversight. Additionally, systems should be positioned as assistive rather than authoritative. Furthermore, dataset imbalance and limited field data can bias models toward controlled conditions, requiring continued domain-diverse collection. Regional bias remains a risk due to geographic variation in disease appearance; mitigation requires local data collection, fine-tuning, and transparent deployment limits. Environmental and computational costs also remain a concern. Although parameter-efficient tuning reduces overhead, architectural choices must balance sustainability with accuracy and robustness. Ultimately, careful design, transparent reporting, and responsible deployment are essential for equitable agricultural benefit.

References

Muhammad Awais, Ali Husain Salem Abdulla Alharthi, Amandeep Kumar, Hisham Cholakkal, and Rao Muhammad Anwer. Agrogpt: Efficient agricultural vision-language model with expert tuning, 2024. URL https://arxiv.org/abs/2410.08405.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Hritik Bansal, Devandra Singh Sachan, Kai-Wei Chang, Aditya Grover, Gargi Ghosh, Wen-tau Yih, and Ramakanth Pasunuru. Honeybee: Data recipes for vision-language reasoners. *arXiv preprint arXiv:2510.12225*, 2025.

Vardhan Dongre, Chi Gui, Shubham Garg, Hooshang Nayyeri, Gokhan Tur, Dilek Hakkani-Tür, and Vikram S Adve. Mirage: A benchmark for multimodal information-seeking and reasoning in agricultural expert-guided conversations. *arXiv preprint arXiv:2506.20100*, 2025.

Jawad Hassan, Kaleem Razzaq Malik, Ghulam Irtaza, Ali Ghulam, and Ashfaq Ahmad. Disease identification using deep learning in agriculture: A case study of cotton plant. *VFAST Transactions on Software Engineering*, 10(4):104–115, Dec. 2022. doi: 10.21015/vtse.v10i4.1224. URL https://vfast.org/journals/index.php/VTSE/article/view/1224.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

David. P. Hughes and Marcel Salathe. An open access repository of images on plant health to enable the development of mobile disease diagnostics, 2016. URL https://arxiv.org/abs/1511.08060.

Md. Manowarul Islam, Md Abdul Ahad Adil, Md. Alamin Talukder, Md. Khabir Uddin Ahamed, Md Ashraf Uddin, Md. Kamran Hasan, Selina Sharmin, Md. Mahbubur Rahman, and Sumon Kumar Debnath. Deepcrop: Deep learning-based crop disease prediction with web application. *Journal of Agriculture and Food Research*, 14:100764, 2023. ISSN 2666-1543. doi: https://doi.org/10.1016/j.jafr.2023.100764. URL https://www.sciencedirect.com/science/article/pii/S2666154323002715.

Sudhesh K.M., Sowmya V., Sainamole Kurian P., and Sikha O.K. Ai based rice leaf disease identification enhanced by dynamic mode decomposition. *Engineering Applications of Artificial Intelligence*, 120:105836, 2023. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2023.105836. URL https://www.sciencedirect.com/science/article/pii/S0952197623000209.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models, 2025. URL https://arxiv.org/abs/2504.05299.

Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 2016. ISSN 1664-462X. doi: 10.3389/fpls.2016.01419. URL https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2016.01419.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. Plantdoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, CoDS COMAD 2020, pp. 249–253. ACM, January 2020. doi: 10.1145/3371158.3371196. URL http://dx.doi.org/10.1145/3371158.3371196.

Ranjita Thapa, Kai Zhang, Noah Snavely, Serge Belongie, and Awais Khan. Plant pathology 2021 - fgvc8. https://kaggle.com/competitions/plant-pathology-2021-fgvc8, 2021. Kaggle.

Tyr Wiesner-Hanks and Mohammed Brahimi. Image set for deep learning: Field images of maize annotated with disease symptoms, Feb 2019. URL osf.io/p67rz.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6757–6767, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.

Yueyue Zhou, Hongping Yan, Kun Ding, Tingting Cai, and Yan Zhang. Few-shot image classification of crop diseases based on vision–language models. *Sensors*, 24(18), 2024. ISSN 1424-8220. doi: 10.3390/s24186109. URL https://www.mdpi.com/1424-8220/24/18/6109.

A Dataset Composition

A.1 Source Datasets

Dataset Name	Source	Crop	# Diseases	# Samples	Citation/Link
New Bangladeshi Crop Disease	Lab	Rice	4	4,078	Moin (2023)
PlantVillage	Lab	Apple	4	4,645	G. & J. (2019)
		Blueberry	1	1,502	
		Cherry	2	2,502	
		Corn	4	4,345	
		Grape	4	4,639	
		Orange	1	5,507	
		Peach	2	3,297	
		Bell Pepper	2	2,478	
		Potato	3	3,000	
		Raspberry	1	1,000	
		Soybean	1	5,090	
		Squash	1	1,835	
		Strawberry	2	2,109	
		Tomato	10	18,835	
PlantDoc	Field	Apple	3	287	Jain & Kayal (2024)
		Bell Pepper	2	125	
		Blueberry	1	117	
		Cherry	1	57	
		Corn	3	378	
		Grape	2	154	
		Peach	1	112	
		Potato	2	379	
		Raspberry	1	119	
		Soybean	1	65	
		Squash	1	130	
		Strawberry	2	96	
		Tomato	9	903	
Apple Dataset 2021	Field	Apple	5	15,675	FGVC (2021)
Roboflow Rice Disease Dataset	Field	Rice	6	3,733	Jain & Kayal (2024)
Paddy Doctor	Field	Rice	10	10,407	Paddy Doctor (2023)
Rice Leaf Disease	Field	Rice	4	5,932	Sethy (2020)
CD&S Dataset	Field	Corn	1	523	Ahmad (2021)
Disease of Maize in the Field	Field	Corn	5	2,716	UoPretoria (2022)
CornNLB	Field	Corn	1	1,787	Brahimi et al. (2018)
Zeytin Olive Leaf Disease	Lab	Olive	3	5,011	Uguz & Uysal (2021)
Strawberry Disease Detection	Field	Strawberry	5	2,268	Afzaal et al. (2021)

Table 3: Datasets used to compile AgriPath-LF16 along with links to the download source

A.2 Downsampling Strategy

Below is the downsampling logic used to create AgriPath-LF16-30k. The balanced subset of AgriPath-LF16 used in this paper:

- **Case 1 (Single Source):** If samples existed for only one source (lab or field), all samples were kept if the count was close to the downsample target (within ± 10), or downsampled to match the target if the count exceeded it by more than 20.
- **Case 2L (Lab less than target, field more):** If the total lab and field samples were less than the target, all samples were kept. Otherwise, field samples were downsampled to the required count, and all lab samples were included.
- **Case 2F (Field less than target, lab more):** Similar to Case 2L, if the total samples were less than the target, all samples were kept. Otherwise, lab samples were downsampled to the required count, and all field samples were included.
- **Case 3 (Both less than target):** For under-represented classes, the crop-disease pair was flagged for review, and all existing samples were kept and added to the dataset for inspection.
- **Case 4 (Both meet or exceed target):** Both lab and field samples were downsampled to equal amounts to achieve a 50/50 split and match the target sample size.

A.3 Dataset Structure

Each sample in AgriPath-LF16 includes the following structured metadata:

- `image`: The image of the crop.
- `crop`: The specific crop (leaf) represented in the image.
- `disease`: The disease present (if any) in the image.
- `source`: Indicates whether the image was taken in a lab or in the field.
- `crop_disease_label`: A combined label representing crop-disease pairs.
- `numeric_label`: The numeric encoding (0-64) for each crop-disease pair.

B Additional CNN Information

All CNN experiments were implemented using PyTorch Lightning. A coarse grid search over batch size and learning rate was conducted for the ResNet-50 model across all experiments. The grid search includes three batch sizes (16, 32, and 64) and three learning rates (1e-4, 2e-4, 5e-4), resulting in a total of 9 experiments. The experiment mentioned in the paper is in bold.

B.1 CNN-Full

Batch Size	Learning Rate	F1(Combined)	F1(Lab)	F1(Field)
16	**1e-4**	**91.0**	**97.2**	**70.3**
16	2e-4	89.3	95.6	66.6
16	5e-4	88.8	94.9	69.9
32	1e-4	90.6	94.0	70.4
32	2e-4	90.1	93.7	70.3
32	5e-4	88.9	91.5	64.5
64	1e-4	89.9	95.0	66.0
64	2e-4	89.3	94.9	64.9
64	5e-4	87.5	91.4	68.6

Table 4: The F1-Scores of the CNN experiments trained on the full dataset (lab and field)

B.2 CNN-Lab

Batch Size	Learning Rate	F1(Combined)	F1(Lab)	F1(Field)
16	**1e-4**	**56.0**	**96.8**	**4.5**
16	2e-4	55.1	94.6	4.9
16	5e-4	55.1	96.8	3.6
32	1e-4	55.5	97.0	4.2
32	2e-4	55.1	95.5	4.7
32	5e-4	55.8	96.3	3.5
64	1e-4	55.4	95.9	4.7
64	2e-4	55.8	96.6	4.2
64	5e-4	54.8	95.3	4.1

Table 5: The F1-Scores of the CNN experiments trained on the lab-only subset of the full dataset

B.3 CNN-Field

Batch Size	Learning Rate	F1 (Combined)	F1(Lab)	F1(Field)
16	**1e-4**	**37.4**	**13.3**	**73.7**
16	2e-4	35.6	10.8	69.6
16	5e-4	37.0	11.6	70.2
32	1e-4	40.2	15.2	76.9
32	2e-4	34.0	8.2	69.1
32	5e-4	34.7	11.2	68.7
64	1e-4	37.3	12.7	73.3
64	2e-4	37.6	12.9	70.6
64	5e-4	36.8	11.4	71.4

Table 6: The F1-Scores of the CNN experiments trained on the field-only subset of the full dataset

C Additional Contrastive VLM Information

C.1 Zero-Shot Templates

Zero-shot classification for contrastive VLMs was performed using prompt-ensemble prototypes constructed from crop–disease metadata. Text prompts were generated using class-specific templates and averaged in embedding space prior to similarity computation. Two template families were defined:

1. Diseased Templates:
   - `"a photo of a {crop} leaf with {disease}"`
   - `"an image of a {crop} leaf affected by {disease}"`
   - `"a close-up photo of a {crop} leaf showing {disease}"`
2. Healthy Templates:
   - `"a photo of a healthy {crop} leaf"`
   - `"an image of a healthy {crop} leaf"`
   - `"a close-up photo of a healthy {crop} leaf"`

For each of the 65 crop–disease classes, all corresponding templates were instantiated, encoded using the model's text encoder, L2-normalised, and averaged to form a class prototype vector. During inference, image embeddings were L2-normalised and cosine similarity was computed against all class prototypes. The predicted label corresponds to the maximum similarity score.

C.2 Linear Probing Configuration

To evaluate the quality of pretrained visual representations, linear probing was performed by training a single fully-connected classifier head on frozen image embeddings.

Backbone Handling:

- Pretrained SigLIP and CLIP backbones were loaded via `AutoModel`.
- All backbone parameters were frozen.
- Image embeddings were extracted using `get_image_features()` when available, or via the CLS token from the vision encoder.
- Embeddings were L2-normalised before classification.

Classifier Head:

A single Linear layer, `Linear(feature_dim, 65)` was trained on top of frozen features.

- Only classifier parameters were updated.
- Backbone gradients were disabled.

The trained head and metadata (feature dimension, class count) were stored as a W&B artifact and reloaded during evaluation.

Training Configuration:

All linear probing experiments used:

- Batch size: 64
- Cross-Entropy loss
- Macro-F1 as evaluation metric
- Learning Rate selected via small manual sweep from {0.001, 0.003, 0.01}

D Additional Generative VLM Information

D.1 Zero-Shot Evaluations

	PSR (%)	F1	Precision	Recall
Qwen2.5-VL-3B-ZS-Pure	90.87	3.9	5.6	5.9
Qwen2.5-VL-3B-ZS-Context	20.02	2.4	18.1	2.0
Qwen2.5-VL-3B-ZS-MCQ	**21.24**	**23.7**	**68.7**	**16.5**
Qwen2.5-VL-7B-ZS-Pure	68.76	2.1	2.8	3.3
Qwen2.5-VL-7B-ZS-Context	90.21	10.5	17.7	15.5
Qwen2.5-VL-7B-ZS-MCQ	**94.86**	**65.9**	**71.0**	**65.9**
SmolVLM-500M-ZS-Pure	0.17	0	0	0
SmolVLM-500M-ZS-Context	**24.99**	**0.2**	**0.4**	**0.6**
SmolVLM-500M-ZS-MCQ	0	0	0	0

Table 7: All ZS experiments for all Generative VLMs, reporting PSR, F1, Precision, and Recall in a range of 0-100. Experiments mentioned in the paper are in bold

D.2 LoRA Conversational Format

```
conversation = [
    {"role": "system",
        "content": [
            {"type": "text", "text": "You are an expert pathologist and need to identify the crop and
                 disease present in an image. If it is a healthy crop, classify it as healthy"}
        ]
    },
    {"role": "user",
    "content": [
            {"type": "text", "text": "Identify the crop and disease in the image."},
            {"type": "image", "image": sample['image']}
        ]
    },
    {"role": "assistant",
    "content": [
            {"type": "text", "text": f"Class: {sample['crop']}\nDisease: {sample['disease']}"}
        ]
    }
]
```

D.3 Sweep Parameters

All sweeps below use $\alpha = 2r$ in the LoRA configurations of the sweep.

D.3.1 Initial Search Space

Parameters	Distribution	Range
Learning Rate	Uniform	5×10^{-5} to 2×10^{-4}
LoRA Rank r	Categorical	32, 64, 128 (Qwen3) | 64, 128 (Qwen7)
Weight Decay	Uniform	0 to 0.1

Table 8: The initial parameters used for sweeps. Qwen3 and Qwen7 refer to the 3B and 7B variants of the Qwen2.5-VL-xB-LoRA experiments respectively.

Early sweeps indicated stable convergence in the region $r \in \{64, 128\}$, hence the shift from {32, 64, 128} to {64, 128} seen above in Table 8.

D.3.2 Refined Search Space

Based on preliminary results, the LoRA rank was fixed and subsequent sweeps were restricted to learning rate and weight decay, using a reduced search range.

Parameters	Distribution	Range
Learning Rate	Uniform	5×10^{-5} to 1.5×10^{-4}
Weight Decay	Uniform	0 to 0.1
LoRA Rank r	Fixed	128

Table 9: The refined parameters used for sweeps after the initial sweeps showed some trends of convergence.

D.3.3 Summary of Sweep Configurations

Sweep Name	Regime	Search Space
Qwen3	Full LoRA	Initial
Qwen3-LAB	Lab LoRA	Initial
Qwen3-FV	Frozen Vision	Refined
Qwen3-FIELD	Field LoRA	Refined
Qwen7	Full LoRA	Initial
Qwen7-LAB	Lab LoRA	Initial
Qwen7-FV	Frozen Vision	Refined
Qwen7-FIELD	Field LoRA	Refined
Smol	Full LoRA	Refined
Smol-LAB	Lab LoRA	Refined
Smol-FV	Frozen Vision	Refined
Smol-FIELD	Field LoRA	Refined

Table 10: A summary of the search spaces used for each sweep. Qwen3 relates to Qwen2.5-VL-3B, Qwen7 relates to Qwen2.5-VL-7B, and Smol relates to SmolVLM-500M. Initial relates to appendix D.3.1, and Refined relates to appendix D.3.2

D.4 Implementation Configuration

D.4.1 Training Frameworks and LoRA Implementation

Generative VLM experiments were conducted using two training backends:

- Qwen2.5-VL models (3B, 7B) were fine-tuned using the UnslothAI framework via `FastVisionModel` with PEFT-based LoRA adaptation.
- SmolVLM-500M was fine-tuned using a custom PEFT pipeline built on `Idefics3ForConditionalGeneration`, as UnslothAI does not natively support this architecture.

In both cases, LoRA was applied to attention and MLP modules. For frozen-vision experiments, LoRA was restricted to language components only, leaving vision layers untouched. All LoRA runs used:

- Scaling Factor: $\alpha = 2r$
- Dropout: 0
- Bias: none

This ensures architectural parity across implementations despite backend differences.

D.4.2 Inference Configuration

During evaluation:

- Models were loaded using their respective backends (UnslothAI or PEFT).
- Generation was deterministic (temperature = 0).
- Image resizing was constrained to a longest edge of 512 pixels.
- No beam search or sampling was used.

This ensured that performance differences reflect representational differences rather than stochastic decoding effects.

D.4.3 Output Parsing and Evaluation

Model outputs were programmatically mapped to the 65 crop–disease classes. The following regular expressions were used to extract structured predictions:

- `"(?:Class|Answer|Crop):\s*(\w+(?: \w+)*)\s*[\r\n]+Disease:\s*(\w+(?:_\w+)*)"`
- `"Answer:\s*[\r\n]+(\w+(?: \w+)*)\s*[\r\n]+(\w+(?:_\w+)*)"`
- `"Disease:\s*(\w+(?:_\w+)*)\s*[\r\n]+(?:Crop|Class|Answer):\s*(\w+(?: \w+)*)"`

Parsing then extracted the exact crop and disease fields from the generated text and matched it against canonical class list. If no valid mapping is found, the prediction is assigned to the `false_parse` class.

Empty generations and malformed outputs were treated as incorrect predictions and penalised F1 computation in order to retrieve a holistic view of model capabilities.

E Confusion Matrix Analysis

E.1 CNN – ResNet-50 (Batch=16, LR=1e-4)

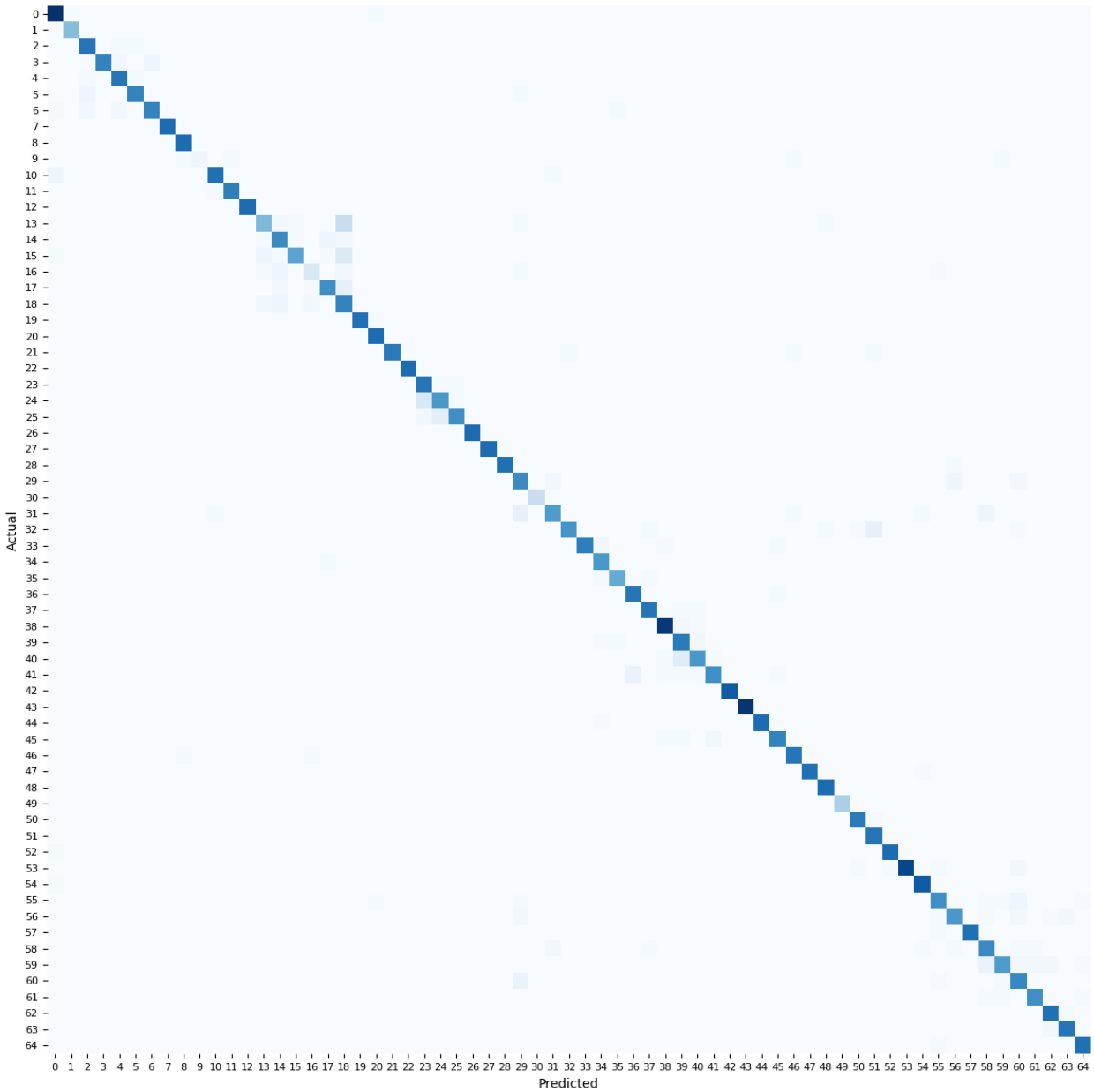


Figure 5: Top confusion pairs are discussed in Section 4.5.1. Actual labels are along the y-axis, and Predicted labels are along the x-axis.

E.2 CLIP – CLIP/L/14

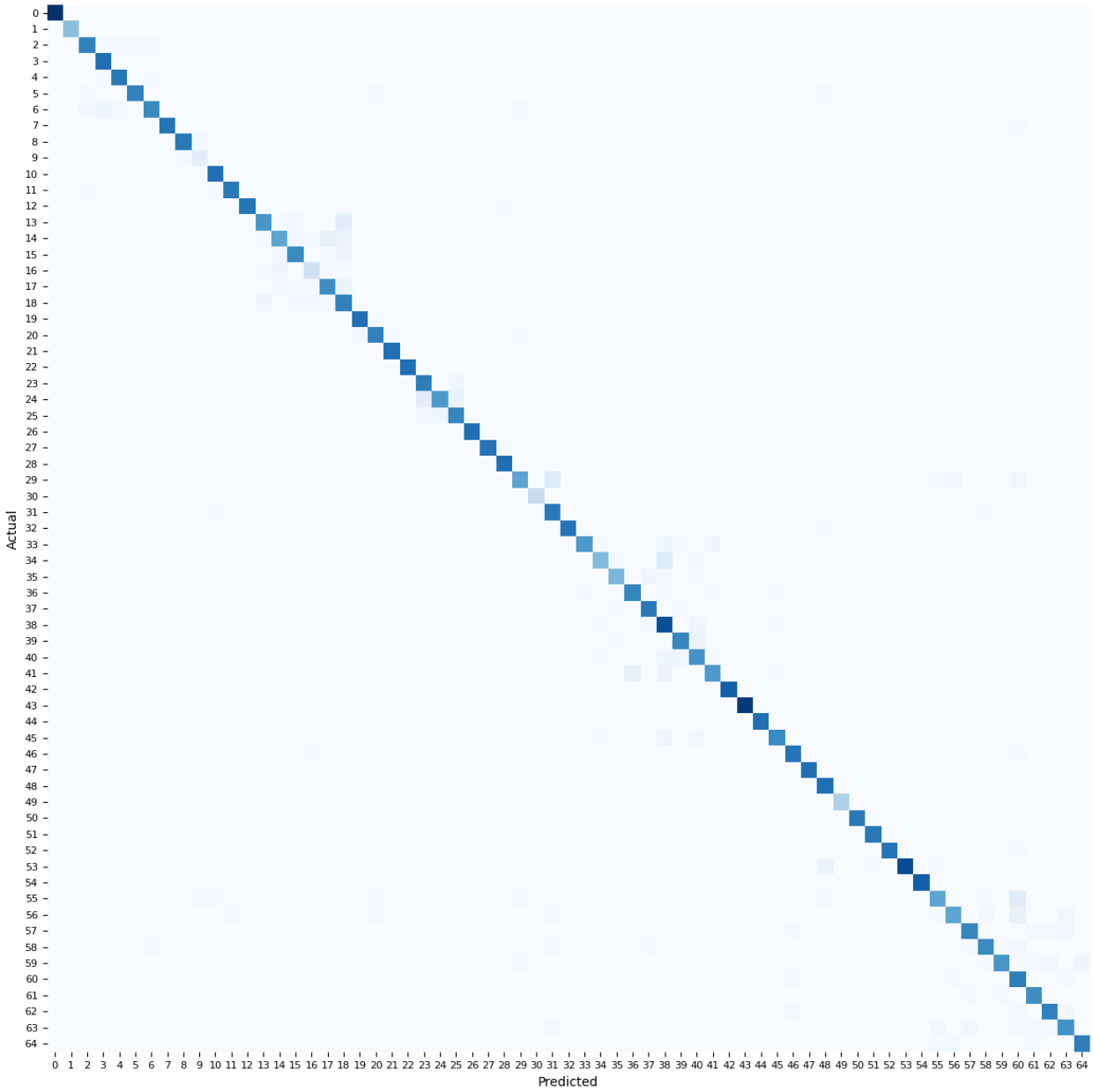


Figure 6: Top confusion pairs are discussed in Section 4.5.2. Actual labels are along the y-axis, and Predicted labels are along the x-axis.