Computers and Electronics in Agriculture 247 (2026) 111692

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

An end-to-end vision–language framework with LLM-based reasoning for decision-oriented tomato leaf disease management in greenhouse environments

Muhammad Shafay [a,*], Divya Velayudhan [a], Muhammad Owais [b,*], Taimur Hassan [c], Lakmal Seneviratne [b], Irfan Hussain [b], Naoufel Werghi [a]

[a] Department of Computer Science, Khalifa University of Science & Technology, Abu Dhabi, United Arab Emirates
[b] Khalifa University Center for Autonomous Robotic Systems (KUCARS), Khalifa University of Science & Technology, Abu Dhabi, United Arab Emirates
[c] Department of Electrical, Computer and Biomedical Engineering, Abu Dhabi University, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Dataset link: GitHub, Figshare

Keywords:
Vision–language models
Contrastive learning
Lesion-aware learning
Agricultural decision support
Image–text alignment

ABSTRACT

Tomato foliar diseases pose a persistent threat to global food security, causing substantial yield losses and relying heavily on subjective, labor-intensive manual inspection. Although deep learning models achieve high accuracy under controlled laboratory conditions, their performance degrades markedly in real agricultural environments due to visual variability, background clutter, and inconsistent illumination. Moreover, existing automated systems fail to reflect how agricultural experts diagnose diseases by jointly considering global plant condition and localized symptom patterns. This paper presents an end-to-end framework for tomato leaf disease management that integrates vision–language disease recognition, lesion-aware representation learning, spatial aggregation, and large language model (LLM)–based reasoning for decision-oriented greenhouse deployment. At the core of the framework is the Dual-Level Contrastive Alignment Framework (DLCAF), which mirrors expert diagnostic reasoning by combining Global Disease-Description Alignment to accommodate linguistic and regional variability with Local Symptom Consistency Learning to focus on disease-specific lesions using expert annotations. Leaf-level predictions are aggregated into row- and greenhouse-level indicators to enable spatial analysis and temporal monitoring, which are subsequently interpreted by an LLM to generate actionable management recommendations. Extensive evaluation across laboratory, field, and greenhouse environments demonstrates the effectiveness of the proposed approach. DLCAF achieves 58.1% accuracy on PlantVillage, 72.5% under natural field conditions on FieldPlant, and 36.2% on the Taiwan Tomato Leaf Disease Dataset, consistently outperforming state-of-the-art vision–language models including CLIP, LongCLIP, and AgriCLIP. Additional validation on a newly collected real-world greenhouse dataset from Abu Dhabi (502 images) further confirms robust performance under practical deployment conditions, achieving 74.3% accuracy. Ablation studies show that expert-guided lesion supervision and multi-description semantic alignment provide more informative learning signals than random sampling or single-caption training. These results demonstrate that integrating global semantic understanding, localized symptom analysis, and LLM-based reasoning enables robust, interpretable, and practically deployable disease detection and decision-support systems for precision agriculture. To support reproducibility and future research, the implementation code is publicly available at GitHub, along with our collected AD Dataset released at Figshare.

1. Introduction

Tomato production plays a critical role in global food security and economic sustainability, with worldwide yields exceeding 192 million tonnes across 4.9 million hectares in 2023 (Food and Agriculture Organization of the United Nations, 2023). Tomatoes also serve as a key raw material for several high-value industries, including food processing, cosmetics, and pharmaceuticals. Despite their broad economic importance, foliar diseases continue to threaten tomato production worldwide, often reducing yields by 20–40% (Panno et al., 2021). Reliable disease identification is therefore essential for maintaining crop

* Corresponding author.
E-mail addresses: 100057573@ku.ac.ae (M. Shafay), divya.velayudhan@ku.ac.ae (D. Velayudhan), muhammad.owais@ku.ac.ae (M. Owais), taimur.hassan@adu.ac.ae (T. Hassan), lakmal.seneviratne@ku.ac.ae (L. Seneviratne), irfan.hussain@ku.ac.ae (I. Hussain), naoufel.werghi@ku.ac.ae (N. Werghi).

https://doi.org/10.1016/j.compag.2026.111692
Received 23 December 2025; Received in revised form 1 March 2026; Accepted 17 March 2026
Available online 26 March 2026

health; however, current agricultural practice relies heavily on manual inspection by experts, a process that is subjective, labor-intensive, and difficult to scale in commercial greenhouse environments.

Recent advances in artificial intelligence have accelerated the development of automated systems for diagnosing plant diseases (Too et al., 2019). Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) achieve high accuracy on controlled datasets, often exceeding 95% on benchmarks such as PlantVillage (George et al., 2025). However, this performance rarely translates to real-world conditions, with accuracy typically dropping by 30–40 percentage points when models trained on laboratory datasets are evaluated on field-level benchmarks such as PlantDoc or FieldPlant (Singh et al., 2020; Moupojou et al., 2023). Factors such as variable lighting, background clutter, occlusions, leaf pose variation, and natural symptom diversity reveal that many models rely on superficial correlations rather than robust disease representations, leading to poor generalization in field environments.

Vision–language models (VLMs) offer a promising alternative by learning joint visual–semantic representations. CLIP established the modern paradigm of scalable vision–language pretraining using large image-text datasets (Radford et al., 2021), while agricultural adaptations such as AgriCLIP and SCOLD incorporate domain-specific symptom descriptions to improve robustness under field conditions (Nawaz et al., 2025; Nguyen Quoc et al., 2026). Despite these advances, existing agricultural VLMs remain limited for fine-grained diagnosis of tomato diseases. Each image is typically paired with a single, full-image caption, restricting the model's ability to capture lesion-specific cues and the linguistic variability associated with disease progression and symptom expression. As a result, these models fail to emphasize the localized pathological patterns that are central to expert diagnosis.

Beyond accurate leaf-level recognition, practical agricultural deployment requires systems that can aggregate observations across space and time to support actionable, crop-level decision-making. This need is particularly acute in greenhouse environments, where high-density cultivation amplifies the impact of delayed or imprecise disease management (Kanwar, 2013). Practitioners require interpretable insights at the greenhouse level, such as spatial disease distribution, row-wise infection trends, and temporal progression, rather than isolated leaf-level predictions. Conventional classifiers and existing VLMs do not natively support such structured aggregation or reasoning.

Large Language Models (LLMs) offer a complementary capability by enabling contextual reasoning over structured inputs, supporting the interpretation of spatial and temporal disease indicators, and generating expert-like management recommendations. Integrating LLMs with vision-based disease recognition therefore provides a pathway to transform raw predictions into actionable decision support.

To address these challenges, we propose the Dual-Level Contrastive Alignment Framework (DLCAF) together with a field-aware post-processing module and an LLM-based advisory system. DLCAF learns disease semantics through Global Disease-Description Alignment (GDDA), which associates each disease with multiple textual descriptions, and Local Symptom Consistency Learning (LSCL), which enforces lesion-level consistency using expert annotations. The post-processing module aggregates DLCAF predictions into row-level and greenhouse-level prevalence metrics, providing interpretable indicators of infection pressure. An LLM then interprets these structured indicators over time to produce actionable guidance, including hotspot identification, intervention prioritization, and monitoring recommendations. This unified DLCAF–LLM framework mirrors expert diagnostic workflows by coupling robust multimodal disease recognition with practical decision support for greenhouse environments.

The main contributions of this work are summarized as follows:

- We introduce DLCAF, a dual-level contrastive learning framework that integrates multi-description semantic alignment with lesion-aware representation learning to enhance robustness under domain shift.
- We propose Local Symptom Consistency Learning (LSCL), a supervised lesion-focused contrastive objective that captures fine-grained pathological invariance beyond full-image supervision.
- We develop a field-aware spatial aggregation module that transforms leaf-level predictions into row- and greenhouse-level prevalence indicators for interpretable disease monitoring.
- We validate the framework across laboratory, field, mixed, and real greenhouse datasets, including a newly collected Abu Dhabi (AD) greenhouse dataset comprising 502 images, demonstrating strong cross-domain generalization.
- We integrate an LLM-based spatiotemporal advisory component that interprets structured greenhouse-level indicators to generate actionable management recommendations.

2. Related work

2.1. Vision-centric plant disease recognition

Recent advances in deep learning have led to substantial improvements in plant disease classification (Shafay et al., 2025). Early convolutional and transformer-based approaches primarily focused on enhancing feature extraction and architectural design. Satyanarayana et al. (2025) developed ECSAN to improve multi-crop disease recognition using Capsule Networks and shuffle attention, effectively preserving lesion geometry and part–whole relationships. Ghosh et al. (2025) proposed a Differential Evolution–driven strategy to optimize Vision Transformer architectures for tomato disease classification, while Sundhar et al. (2025) introduced a GAT-GCN hybrid that models superpixel segments as graph nodes to capture relational patterns between lesions. Additional studies explored CNN–ML hybrids for robustness under noisy conditions (Khan et al., 2024), lightweight edge-deployed classifiers for real-time detection (Karim et al., 2024), and capsule-based architectures for modeling spatial lesion structure (Abouelmagd et al., 2024).

Although these methods achieve strong performance on benchmark datasets, they operate primarily at the image level and rely solely on visual supervision. As a result, they often struggle to generalize from controlled laboratory conditions to complex field or greenhouse environments, where illumination variability, background clutter, and occlusion introduce significant domain shift. Moreover, purely vision-centric models lack semantic grounding and do not naturally support structured reasoning beyond leaf-level classification.

2.2. Vision–language models and cross-modal alignment

Vision–language models (VLMs) extend vision-only approaches by learning joint visual–semantic representations through cross-modal contrastive alignment. CLIP-style training establishes a shared embedding space between images and textual descriptions, enabling zero-shot transfer across tasks and domains. Building on this paradigm, AgriCLIP (Nawaz et al., 2025) adapts CLIP to agricultural imagery using domain-specialized image–text alignment, while SCOLD (Nguyen Quoc et al., 2026) introduces soft-target contrastive learning to enhance robustness under ambiguous symptoms and long-tailed distributions.

In addition, recent work has explored automatically generating disease descriptions from images to enhance multimodal fusion. A multimodal framework based on large multimodal models (e.g., LLaVA and CogAgent) was proposed to generate structured textual descriptions from crop images and fuse them with visual features through a Projected Visual–Textual Discriminant (PVD) module (Xiong et al., 2025). While this approach improves cross-modal integration by leveraging automatically generated text, it remains primarily focused on image-level multimodal fusion and does not incorporate lesion-level supervision or structured spatial reasoning.

Beyond language-based alignment, multimodal fusion has also been explored through the integration of image data and environmental

Table 1
Comparison of recent plant leaf disease recognition frameworks in terms of modeling scope and system-level capabilities. ✓ = supported; ✗ = not addressed.

Authors, Venue, Year	Disease Ident.	Cross-Data Eval.	Vision–Language Modeling			Spatiotemporal	
			Text Align	Multi Desc	Lesion Level	Temp. Analysis	LLM Supp.
Khan *et al.*, *Sci. Rep.*, 2024 (Khan et al., 2024)	✓	✗	✗	✗	✗	✗	✗
Satyanarayana *et al.*, *Comp. El. Agri.*, 2025 (Satyanarayana et al., 2025)	✓	✗	✗	✗	✗	✗	✗
Ghosh *et al.*, *Comp. El. Agri.*, 2025 (Ghosh et al., 2025)	✓	✗	✗	✗	✗	✗	✗
Nawaz *et al.*, *COLING*, 2024 (AgriCLIP) (Nawaz et al., 2025)	✓	✓	✓	✗	✗	✗	✗
Quoc *et al.*, *Expert Syst. Appl.*, 2025 (SCOLD) (Nguyen Quoc et al., 2026)	✓	✓	✓	✗	✗	✗	✗
Yan *et al.*, *Comp. El. Agri.*, 2025 (Yan et al., 2025)	✓	✗	✗	✗	✗	✗	✓
Zhu *et al.*, *Comp. El. Agri.*, 2025 (Zhu et al., 2025)	✓	✗	✗	✗	✗	✗	✓
Li *et al.*, *Comp. El. Agri.*, 2026 (Li et al., 2026)	✓	✗	✗	✗	✗	✗	✓
Huang *et al.*, *Comp. El. Agri.*, 2026 (Huang et al., 2026)	✓	✗	✓	✗	✗	✗	✓
DLCAF + LLM (Ours)	✓	✓	✓	✓	✓	✓	✓

sensor information. A lightweight multimodal parallel transformer was proposed for apple disease segmentation and severity classification by combining visual features with multi-dimensional sensor inputs via cross-scale attention mechanisms (Song et al., 2025). Although effective for sensor-level fusion and efficient deployment, this line of work does not explore semantic alignment with textual descriptions or language-driven advisory reasoning. Overall, while cross-modal alignment improves semantic grounding relative to purely visual classifiers, most existing VLM-based agricultural systems remain centered on per-image recognition. They do not explicitly incorporate lesion-level modeling or structured aggregation mechanisms that reflect expert diagnostic workflows.

2.3. LLM-integrated and multimodal agricultural systems

Recent studies have further investigated multimodal question-answering and advisory systems for agriculture. A joint topic entity and intent recognition model based on a multimodal Transformer was proposed to handle image–question pairs in agricultural disease and pest QA (Huang et al., 2026). By employing a dual-tower architecture with cross-modal interaction and multi-task learning, the model jointly recognizes user intent and disease entities. Although effective for multimodal intent understanding, the framework focuses on question interpretation rather than structured disease monitoring or spatiotemporal aggregation.

Parallel research has explored integrating visual models with large language models (LLMs) to move toward decision-support systems. CDIP-ChatGLM3 combines EfficientNet-B2 with a fine-tuned ChatGLM3 model to provide disease identification and prescription generation (Yan et al., 2025). PotatoGPT integrates multimodal disease detection with GPT-4 to enable online diagnosis and prevention recommendations (Zhu et al., 2025). KiwiGuard proposes a collaborative visual–LLM architecture enhanced with retrieval-augmented generation (RAG) and multimodal knowledge graphs to mitigate hallucination (Li et al., 2026), while WDLM incorporates reasoning chains and segmentation mechanisms for wheat disease diagnosis (Zhang et al., 2024). While these systems combine perception and language reasoning, they typically employ loosely coupled dual-model pipelines, limiting multimodal interaction to per-image interpretation without lesion-level modeling or hierarchical spatial aggregation.

Collectively, these studies demonstrate the growing interest in multimodal agricultural intelligence; however, existing approaches typically emphasize either cross-modal alignment at the image level or loosely coupled vision–LLM pipelines without unified lesion-aware modeling and hierarchical spatial reasoning. In contrast, DLCAF+LLM unifies multi-description semantic alignment, supervised lesion-level contrastive learning, hierarchical spatial aggregation, and LLM-driven spatiotemporal advisory generation within a single end-to-end framework. By explicitly modeling pathological structure, semantic variability, and greenhouse-level infection dynamics, the proposed approach enables robust cross-domain recognition and interpretable, decision-oriented disease management. Table 1 summarizes the differences in modeling scope and system-level capabilities across existing approaches. The remainder of this paper is organized as follows. Section 3 introduces the proposed methodology, which integrates the Dual-Level Contrastive Alignment Framework (Section 3.1) with LLM-based spatiotemporal reasoning (Section 3.2). Section 4 details the experimental setup and dataset characteristics, while Section 5 presents the quantitative results and ablation studies. Section 6 provides a comprehensive analysis of the findings, focusing on domain shift and practical implications for agricultural deployment. Finally, Section 7 summarizes the key contributions of this work and outlines directions for future research.

3. Proposed methodology

We propose an end-to-end framework that integrates vision–language disease recognition, spatial aggregation, and large language model (LLM)– based reasoning to support informed and actionable decision-making in greenhouse environments. As illustrated in Fig. 1, the pipeline begins with leaf images collected during multiple inspection cycles (at times t_0 and t_n), capturing the evolving health status of the crop. Each image is processed by the proposed Dual-Level Contrastive Alignment Framework (DLCAF), which combines global disease description alignment with local symptom consistency learning to produce leaf-level disease predictions. These predictions are subsequently mapped to their corresponding greenhouse rows and categorized as healthy or infected.

To derive greenhouse-level insights, leaf-level predictions are aggregated across rows by computing row-wise infection percentages, summarizing disease-category distributions, and identifying rows with elevated disease pressure (referred to as *hotspots*). These aggregated metrics provide a concise representation of disease burden and spatial heterogeneity at each inspection time. For higher-level reasoning, the aggregated spatial descriptors are encoded into a fixed-format prompt that contains key numerical indicators, including the overall infection rate, disease-wise prevalence, and row-level infection values. When multiple inspection cycles are available, these summaries are temporally ordered to enable trend analysis and comparison across time.

The structured prompt is then provided to an LLM (Llama 3), which organizes the spatial and temporal indicators into a consistent representation to facilitate reliable interpretation across inspection cycles. Using this information, the LLM performs contextual reasoning to generate a greenhouse-level disease summary, identify and interpret hotspot patterns, compare conditions between t_0 and t_n, and recommend appropriate management actions. These recommendations include prioritization of interventions, treatment guidance, and suggested monitoring intervals. By integrating DLCAF's discriminative capabilities with spatial aggregation and LLM-based inference, the proposed framework delivers an interpretable and practical solution for precision disease management in controlled-environment agriculture.

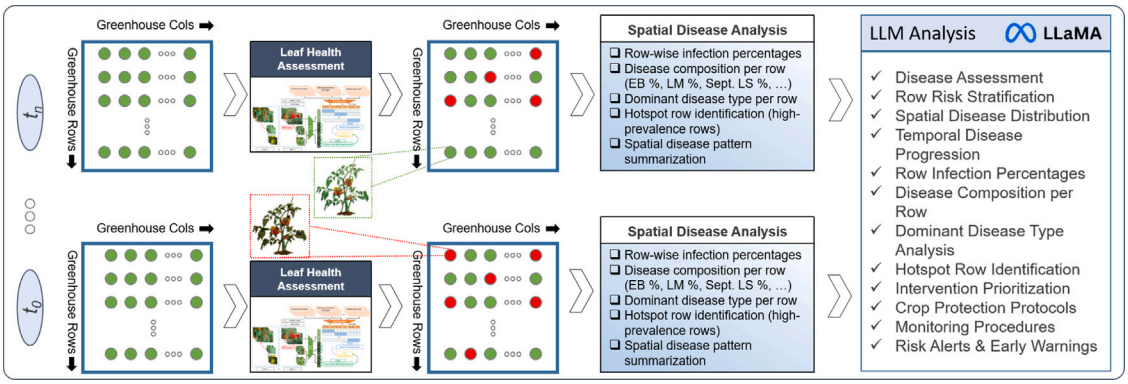

Fig. 1. Overview of the proposed DLCAF–LLM pipeline for spatial tomato disease analysis. DLCAF processes leaf images collected at inspection times t_0 and t_n to generate leaf-level disease predictions across greenhouse rows. These predictions are aggregated into row-level and greenhouse-level spatial descriptors and passed to an LLM (Llama 3), which generates high-level insights including hotspot identification, spatial spread interpretation, temporal comparison, and actionable treatment recommendations.

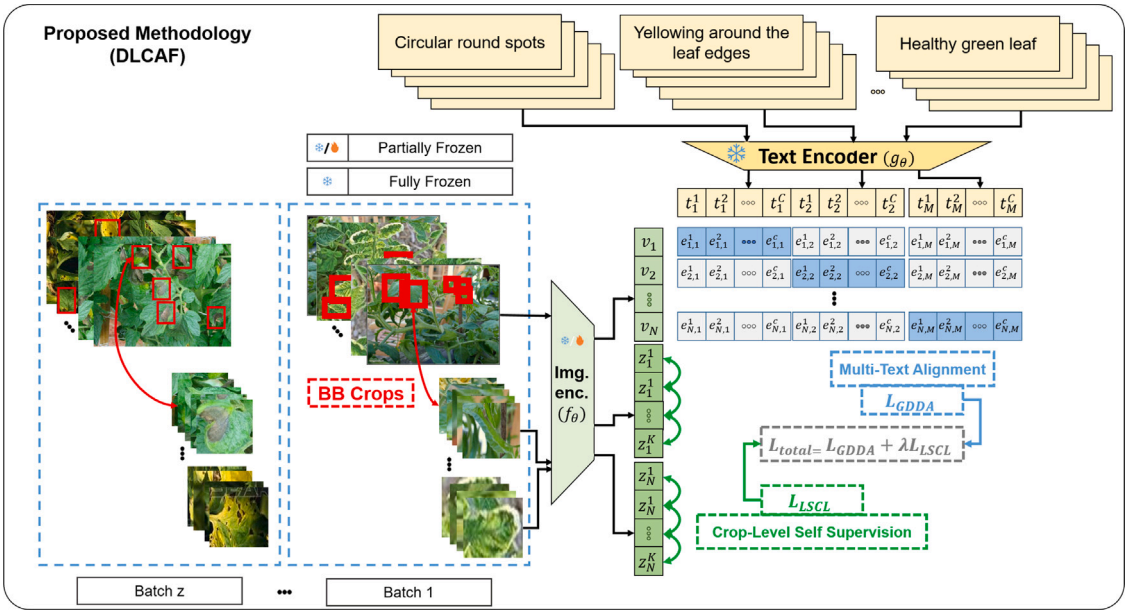

Fig. 2. Overview of DLCAF. Each image $\mathbf{I}_i$ is represented by a global embedding v_i and K lesion-level crops. GDDA performs multi-positive contrastive learning by aligning v_i with multiple disease descriptions $\{t_y^k\}_{k=1}^M$ from a frozen text encoder. Simultaneously, LSCL applies supervised contrastive learning to crop embeddings to enforce intra-class consistency. The joint objective $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GDDA}} + \lambda \mathcal{L}_{\text{LSCL}}$ bridges global semantic alignment with fine-grained pathological discrimination.

3.1. Dual-Level Contrastive Alignment Framework (DLCAF)

We propose the Dual-Level Contrastive Alignment Framework (DLCAF) for tomato leaf disease identification, which fundamentally transforms multimodal supervision in vision–language models. By transcending conventional contrastive approaches restricted to single-instance augmentations, DLCAF introduces a hierarchical learning paradigm that simultaneously captures disease semantics through diverse linguistic formulations and enforces pathological consistency across expert-annotated lesion regions. Fig. 2 shows the training methodology of the proposed framework.

3.1.1. Problem formulation

Traditional supervised disease recognition formulates classification as a discrete label prediction task over a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where each image $\mathbf{x}_i \in \mathbb{R}^{H\times W\times 3}$ is assigned a categorical label $y_i \in \{1, \ldots, C\}$. While effective for categorization, this formulation collapses rich pathological semantics into discrete labels and ignores the diversity of linguistic descriptions and spatial lesion patterns. Vision–language models instead frame recognition as semantic alignment in a shared embedding space. Given paired image–text data $\{(\mathbf{I}_i, t_i)\}_{i=1}^N$, visual and textual encoders $f_\theta(\cdot)$ and $g_\phi(\cdot)$ map inputs into a joint embedding space $\mathbb{R}^d$:

$$v_i = \frac{f_\theta(\mathbf{I}_i)}{\|f_\theta(\mathbf{I}_i)\|_2}, \qquad t_i = \frac{g_\phi(t_i)}{\|g_\phi(t_i)\|_2}. \tag{1}$$

Similarity is measured via cosine similarity as $\text{sim}(v_i, t_j) = v_i^\top t_j$. Following contrastive learning principles, temperature scaling and exponentiation define:

$$e_{i,j} = \exp\left(\frac{v_i^\top t_j}{\tau}\right), \tag{2}$$

where $\tau > 0$ controls distribution sharpness.

Under the standard CLIP formulation, each image in the mini-batch is paired with a single corresponding textual description. Let a mini-batch contain N image–text pairs $\{(I_i, t_i)\}_{i=1}^N$. After computing temperature-scaled exponentiated similarities $e_{i,j}$ as defined in Eq. (2), contrastive learning is applied using the InfoNCE objective. For the image-to-text direction, each image embedding v_i is treated as an

anchor, its paired text t_i as the positive sample, and all other texts in the batch $\{t_j\}_{j\neq i}$ as negatives. The resulting loss is

$$\mathcal{L}_{\mathrm{I}\to\mathrm{T}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e_{i,i}}{\sum_{j=1}^{N} e_{i,j}}, \tag{3}$$

where the numerator corresponds to the similarity between a matched image–text pair, and the denominator normalizes over similarities between the image and all texts in the batch. This objective encourages the embedding of each image to be closer to its paired text than to any other textual description. Similarly, in the text-to-image direction, each text embedding is treated as an anchor, and its paired image as the positive sample. The corresponding loss is

$$\mathcal{L}_{\mathrm{T}\to\mathrm{I}} = -\frac{1}{N}\sum_{j=1}^{N}\log\frac{e_{j,j}}{\sum_{i=1}^{N} e_{i,j}}. \tag{4}$$

The two directional losses are typically combined to form a symmetric contrastive objective that aligns both modalities simultaneously. The full CLIP objective is symmetric:

$$\mathcal{L}_{\mathrm{CLIP}} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{I}\to\mathrm{T}} + \mathcal{L}_{\mathrm{T}\to\mathrm{I}}\right). \tag{5}$$

In our setting, the text encoder $g_\phi(\cdot)$ is frozen to preserve pretrained linguistic knowledge. Therefore, gradients propagate only through the visual encoder, and we optimize solely the image-to-text alignment term $\mathcal{L}_{\mathrm{I}\to\mathrm{T}}$.

3.1.2. Global Disease-Description Alignment (GDDA)

Standard CLIP enforces strict one-to-one alignment between each image and a single textual description. However, in real-world pathology, a disease rarely admits a unique linguistic expression. A single disease may be described through variations in morphology (e.g., circular lesions), color transitions (e.g., yellowing margins), or progression stages (e.g., early necrosis). Restricting supervision to a single sentence per image therefore underutilizes the semantic richness available in expert knowledge. To explicitly model this linguistic diversity, we associate each disease class y with a set of M semantically valid descriptions: $\{t_y^k\}_{k=1}^{M}$. Each description captures complementary pathological cues while remaining consistent with the same underlying disease. Let the normalized textual embedding is defined as:

$$t_y^k = \frac{g_\phi(t_y^k)}{\|g_\phi(t_y^k)\|_2}, \tag{6}$$

Further the temperature-scaled exponentiated similarity between image I_i and description t_y^k is defined as:

$$e_{i,y}^k = \exp\left(\frac{v_i^\top t_y^k}{\tau_g}\right). \tag{7}$$

For an image $\mathbf{I}_i$ with ground-truth label y_i, all descriptions belonging to class y_i constitute valid semantic positives as $\mathcal{P}_i = \{t_{y_i}^k \mid k = 1,\ldots,M\}$. Unlike the standard CLIP loss, which selects a single positive term, GDDA treats the entire set $\mathcal{P}_i$ as positive supervision. Instead of using only one similarity score, GDDA aggregates contributions from all valid descriptions: $\mathrm{Pos}(i) = \sum_{k=1}^{M} e_{i,y_i}^k$. This multi-positive aggregation allows the image embedding to align with a *semantic manifold* of descriptions rather than a single linguistic instance. Geometrically, this encourages the image embedding to move toward the subspace spanned by all valid textual representations of the same disease. The normalization term sums over all classes and their descriptions as $\mathrm{All}(i) = \sum_{y=1}^{C}\sum_{k=1}^{M} e_{i,y}^k$. This ensures that the probability mass assigned to the positive set is measured relative to the entire textual space. To maintain scale consistency with the original InfoNCE formulation and to prevent bias toward classes with more descriptions, we normalize by the number of positives M. The resulting GDDA objective is:

$$\mathcal{L}_{\mathrm{GDDA}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\sum_{k=1}^{M} e_{i,y_i}^k}{M\sum_{y=1}^{C}\sum_{k=1}^{M} e_{i,y}^k}. \tag{8}$$

Importantly, we aggregate exponentiated similarities rather than averaging raw similarity scores. Contrastive learning operates in the softmax space, where probabilities are proportional to exponentiated similarities. Summing $\exp(\cdot)$ terms preserves the probabilistic interpretation of InfoNCE as maximizing the likelihood of the correct class under a categorical distribution. This formulation allows any valid description to strongly support alignment, rather than diluting strong semantic matches through linear averaging. The normalization by M maintains scale consistency with the single-positive CLIP objective and prevents bias toward classes with more descriptions. When $M = 1$, the positive set collapses to a single description and Eq. (8) reduces exactly to the standard one-to-one CLIP loss in Eq. (3). GDDA therefore generalizes instance-level alignment to class-level multi-positive semantic alignment, allowing the model to learn disease representations that are robust to linguistic variability and expressive of diverse pathological manifestations.

3.1.3. Local Symptom Consistency Learning (LSCL)

While GDDA captures global semantic alignment between full-image representations and disease descriptions, reliable diagnosis also requires modeling fine-grained pathological structures that manifest locally within symptomatic regions. Many plant diseases are characterized by specific lesion patterns: circular necrotic spots, chlorotic margins, or mosaic textures that may occupy only a small portion of the leaf surface. Global alignment alone may overlook such localized cues. To address this limitation, we introduce *Local Symptom Consistency Learning* (LSCL), a supervised contrastive objective operating at the crop level.

For each image I_i, we extract K lesion-focused crops as $C_i = \{c_i^1,\ldots,c_i^K\}$. When expert-annotated bounding boxes are available, we prioritize these pathologically meaningful regions. If fewer than K annotated regions exist, the remaining crops are sampled from the full image to maintain a fixed batch structure. This hybrid strategy ensures that LSCL leverages expert knowledge while preserving computational regularity. Each crop is encoded using the shared visual backbone and ℓ_2-normalized:

$$\mathbf{z}_i^k = \frac{f_\theta(c_i^k)}{\|f_\theta(c_i^k)\|_2}, \qquad i = 1,\ldots,N,\ k = 1,\ldots,K. \tag{9}$$

The normalization ensures that similarity comparisons are governed by angular distance in the embedding space. To establish a comprehensive representation space for contrastive supervision, we define the full collection of crop embeddings within the current mini-batch as : $\mathcal{Z} = \bigcup_{i=1}^{N}\{\mathbf{z}_i^k\}_{k=1}^{K}$. The $\mathcal{Z}$ denote the set of all NK crop embeddings in the mini-batch.

For an anchor crop $\mathbf{z}_i^k$ with disease label y_i, we define the positive set as all other crops in the batch that share the same disease label: $\mathcal{P}_{i,k} = \{\mathbf{z}_j^r \in \mathcal{Z} \mid y_j = y_i,\ (j,r) \neq (i,k)\}$. Thus, positives correspond to lesion regions from different images that exhibit the same pathological category, while crops from other classes act as negatives. Unlike self-supervised augmentation-based contrastive learning, LSCL leverages disease annotations to construct semantically meaningful positive pairs. We define the temperature-scaled exponential similarity as:

$$q(\mathbf{z}_a,\mathbf{z}_b) = \exp\left(\frac{\mathbf{z}_a^\top\mathbf{z}_b}{\tau_l}\right), \tag{10}$$

where $\tau_l > 0$ controls concentration of similarity scores. The LSCL loss encourages each anchor crop to be closer to all same-class crops than to any other crop in the batch:

$$\mathcal{L}_{\mathrm{LSCL}} = -\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{1}{|\mathcal{P}_{i,k}|}\sum_{\mathbf{z}_p\in\mathcal{P}_{i,k}}\log\frac{q(\mathbf{z}_i^k,\mathbf{z}_p)}{\sum_{\mathbf{z}_b\in\mathcal{Z}\setminus\{\mathbf{z}_i^k\}} q(\mathbf{z}_i^k,\mathbf{z}_b)}. \tag{11}$$

The numerator aggregates similarities between the anchor and all same-class crops, reinforcing intra-class consistency across different images and environmental conditions. The denominator normalizes over all

other crops in the batch, promoting separation between distinct disease categories. The normalization factor $1/|\mathcal{P}_{i,k}|$ balances contributions across anchors with varying numbers of positives. Geometrically, LSCL encourages crop embeddings from the same disease class to form compact clusters in the feature space while pushing apart crops from different diseases. This pathology-aware contrastive mechanism provides fine-grained supervision complementary to the global semantic alignment enforced by GDDA.

3.1.4. Joint cross-modal and cross-scale optimization

The effectiveness of DLCAF stems from the complementary interaction between global semantic alignment and local pathological consistency. GDDA operates at the image level, aligning each disease image with multiple semantically valid textual descriptions to structure the embedding space according to high-level disease semantics. In contrast, LSCL operates at the crop level, enforcing intra-class compactness and inter-class separation among lesion representations. Although these objectives act at different representational scales, they share a common visual backbone $f_\theta(\cdot)$ and are optimized jointly. The combined training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GDDA}} + \lambda \mathcal{L}_{\text{LSCL}}, \tag{12}$$

where λ balances global semantic supervision and local lesion-level discrimination. Through shared gradient flow, GDDA encourages image embeddings to align with the semantic manifold of disease descriptions, while LSCL simultaneously refines feature representations to capture consistent lesion-level characteristics across different instances of the same pathology. This cross-modal (image–text) and cross-scale (image–crop) coupling enables the model to learn representations that are both semantically grounded and pathologically discriminative. By integrating linguistic supervision, lesion localization, and disease-level annotations within a unified optimization framework, DLCAF produces feature representations that are robust, interpretable, and better suited for real-world disease recognition.

3.1.5. Theoretical perspective

DLCAF extends standard contrastive learning along two orthogonal dimensions. First, GDDA generalizes instance-level one-to-one alignment to class-level multi-positive alignment, enabling semantic manifold learning rather than single-caption matching. Second, LSCL introduces supervised crop-level contrastive learning that enforces intra-class compactness and inter-class separability at the lesion scale. By jointly optimizing global cross-modal alignment and local intra-modal discrimination within a shared backbone, DLCAF performs cross-modal and cross-scale representation learning. This dual-level formulation overcomes the limitations of single-scale contrastive frameworks that operate exclusively at either image-level semantics or local feature invariance.

3.2. LLM-driven spatiotemporal disease analysis

3.2.1. Spatial aggregation and Greenhouse-level summarization

Following leaf-level inference, disease predictions are aggregated spatially to derive row- and greenhouse-level indicators. For each greenhouse row r, let N_r denote the number of sampled leaves and N_r^{inf} the number classified as infected. The row-wise infection percentage is computed as $P_r = \frac{N_r^{\text{inf}}}{N_r} \times 100$. Row-level disease composition is quantified by the proportion of samples assigned to each disease category:

$$P_{r,D_k} = \frac{\#\{i : y_i = D_k \ \wedge \ r(i) = r\}}{N_r} \times 100, \quad k = 1, \ldots, K. \tag{13}$$

Rows exhibiting elevated values of P_r or atypical shifts in $\{P_{r,D_k}\}$ relative to neighboring rows are identified as *hotspots*, corresponding to localized outbreak regions or areas of increased disease pressure.

At the greenhouse scale, let $\mathcal{R}$ denote the set of evaluated rows. The overall infection percentage is computed by aggregating leaf counts across all rows as $P_{\text{GH}} = \frac{\sum_{r\in\mathcal{R}} N_r^{\text{inf}}}{\sum_{r\in\mathcal{R}} N_r} \times 100$. Greenhouse-level disease composition is obtained by aggregating class-specific counts:

$$P_{D_k}^{\text{GH}} = \frac{\sum_{r\in\mathcal{R}} \#\{i : y_i = D_k \wedge r(i) = r\}}{\sum_{r\in\mathcal{R}} N_r} \times 100. \tag{14}$$

Together, these descriptors characterize greenhouse-scale infection pressure, relative disease abundance, and spatial heterogeneity, forming a compact summary for downstream reasoning.

3.2.2. Spatiotemporal encoding and LLM-based advisory generation

Each inspection cycle is encoded as a structured feature vector $\mathbf{S}^{(t)} = \left[P_{\text{GH}}^{(t)}, P_{D_1}^{(t)}, \ldots, P_{D_K}^{(t)}, P_1^{(t)}, \ldots, P_N^{(t)} \right]$, which captures greenhouse-level statistics, disease-category prevalences, and row-wise infection values. When multiple inspection times are available, such as a current inspection at time t_n and a prior inspection at time t_0, the vectors are combined into a spatiotemporal representation:

$$\mathbf{S}_{\text{LLM}} = \begin{bmatrix} \text{Date}^{(t_n)} & \mathbf{S}^{(t_n)} \\ \text{Date}^{(t_0)} & \mathbf{S}^{(t_0)} \end{bmatrix}. \tag{15}$$

This representation encodes temporal changes in infection intensity, shifts in disease composition, and the evolution of row-level hotspots. The structured descriptors are embedded into a fixed prompt format and provided to a large language model (LLM) for high-level reasoning.

Using these inputs, the LLM synthesizes greenhouse-wide disease status, identifies abnormal spatial patterns, compares inspection cycles to assess progression or improvement, and generates actionable management recommendations. The output consists of a structured advisory comprising a spatiotemporal disease summary, a risk assessment, and a targeted action plan. By converting quantitative indicators into interpretable insights, the LLM extends automated disease recognition into a scalable decision-support system for greenhouse management. Fig. 3 illustrates the proposed LLM-driven spatiotemporal analysis pipeline and the resulting structured advisory output, which includes a greenhouse-level disease summary, risk assessment, and targeted action plan.

4. Experimental setup

4.1. Datasets

We adopt a comprehensive dataset strategy to evaluate the cross-domain generalization capability of the proposed framework. To ensure domain and class consistency, only tomato leaf subsets are used across all datasets. Training is performed on PlantDoc and CroppedPD, comprising 3688 tomato leaf images across eight disease classes (Singh et al., 2020). Testing evaluation is conducted on four distinct datasets representing a spectrum of imaging conditions, as summarized in Table 2. (i) **PlantVillage** provides controlled-laboratory images with uniform backgrounds (Mohanty et al., 2016); (ii) **FieldPlant** captures natural field scenes with complex environmental noise (Moupojou et al., 2023); (iii) **TTLD** comprises a mixture of laboratory- and field-acquired images (Chang, 2020); and (iv) **AD Dataset (our)**, a self-collected dataset from a commercial greenhouse in Abu Dhabi (Figshare). This local dataset serves as a critical bridge between controlled environments and unstructured field settings, providing 502 high-resolution images captured under varying greenhouse lighting and occlusion conditions.

4.1.1. AD Dataet (our): Abu Dhabi Greenhouse Dataset

To assess the practical deployment of the proposed framework under real greenhouse conditions, we constructed the Abu Dhabi (AD) Greenhouse Dataset, consisting of 502 high-resolution tomato leaf images collected from a commercial greenhouse in Abu Dhabi, United Arab Emirates, in 2025. The dataset was designed to evaluate the cross-domain generalization of vision-based disease recognition models

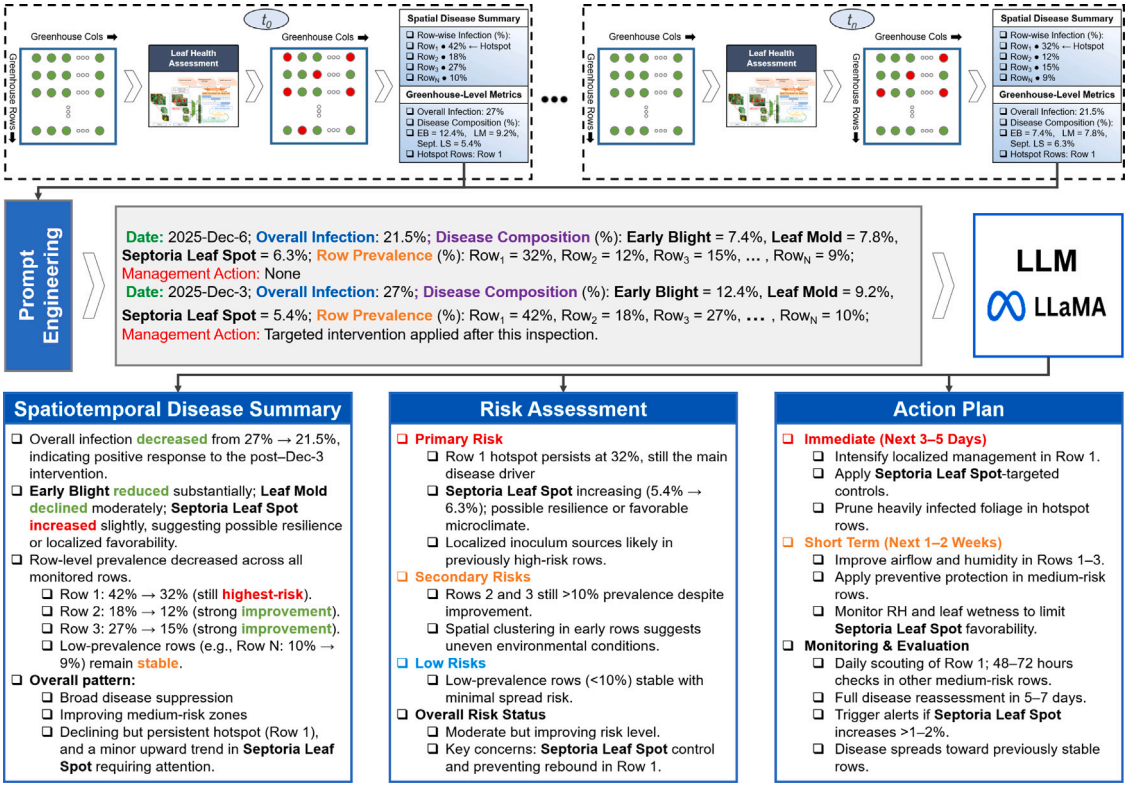

Fig. 3. Overview of the LLM-driven spatiotemporal disease analysis framework. Leaf images collected across greenhouse rows at multiple inspection times are processed by DLCAF to obtain leaf-level disease predictions. These predictions are aggregated into row- and greenhouse-level spatial summaries, encoded into a structured spatiotemporal prompt, and interpreted by an LLM to generate an actionable advisory including disease status, risk assessment, and management recommendations.

Table 2
Dataset composition used for training and evaluation. The AD Dataset represents our local contribution for greenhouse-specific validation.

Dataset	Classes	Total images	Environment	Source
Training Datasets				
PlantDoc	8	758	In-the-wild	Singh et al. (2020)
CroppedPD	8	2930	Cropped/Local	Singh et al. (2020)
Evaluation Datasets				
PlantVillage	10	11,037	Lab	Mohanty et al. (2016)
FieldPlant	5	1204	Field	Moupojou et al. (2023)
TTLD	6	622	Mixed	Chang (2020)
AD Dataset	4	**502**	Greenhouse	**Ours (Local)**

in arid-region greenhouse environments. The AD Dataset is publicly available via Figshare. The dataset includes four classes: Bacterial Spot, Early Blight, Healthy, and Mosaic Virus. Disease labels were assigned based on visual inspection and verified in consultation with an agricultural expert at the greenhouse to ensure consistency with standard symptom descriptions. Images were collected in a structured spatial layout, enabling evaluation of leaf-level disease classification as well as greenhouse-specific visual conditions. Image acquisition was performed using a Samsung Galaxy S23 Ultra smartphone, equipped with a 1/1.3-inch CMOS sensor (Samsung ISOCELL HP2). Images were captured at a resolution of 12.5 megapixels using the wide-angle lens (24 mm equivalent, f/1.7 aperture) with phase-detection autofocus. The dataset reflects realistic greenhouse imaging conditions, including variations in illumination, leaf orientation, and partial occlusion. Representative samples from the AD Dataset are shown in Fig. 4.

4.2. Training and evaluation protocol

We employ a parameter-efficient fine-tuning strategy to adapt pretrained vision–language representations to tomato disease recognition. The text encoder g_ϕ is kept fully frozen to preserve semantic knowledge acquired during large-scale pretraining (Radford et al., 2021). For the vision encoder f_θ, all transformer blocks except the final block are frozen, while the projection head is trainable. This design preserves general visual priors while allowing specialization for disease-specific visual cues. Regarding dataset size, we emphasize that DLCAF does not involve training a large model from scratch. Instead, it adopts a parameter-efficient fine-tuning strategy built upon pretrained CLIP weights, which were originally trained on hundreds of millions of image–text pairs. The text encoder remains fully frozen, and only the final transformer block and projection head of the vision encoder are adapted. Consequently, only a small subset of parameters is updated for tomato disease specialization, substantially reducing data requirements compared to full end-to-end large-scale training. Empirically, the training quantity proves sufficient for adaptation rather than foundational representation learning, as evidenced by strong zero-shot cross-dataset generalization across FieldPlant, TTLD, PlantVillage, and the AD dataset. The stable convergence behavior shown in Fig. 5 further indicates that the model does not exhibit overfitting or instability under the given training regime. Training is performed using the AdamW

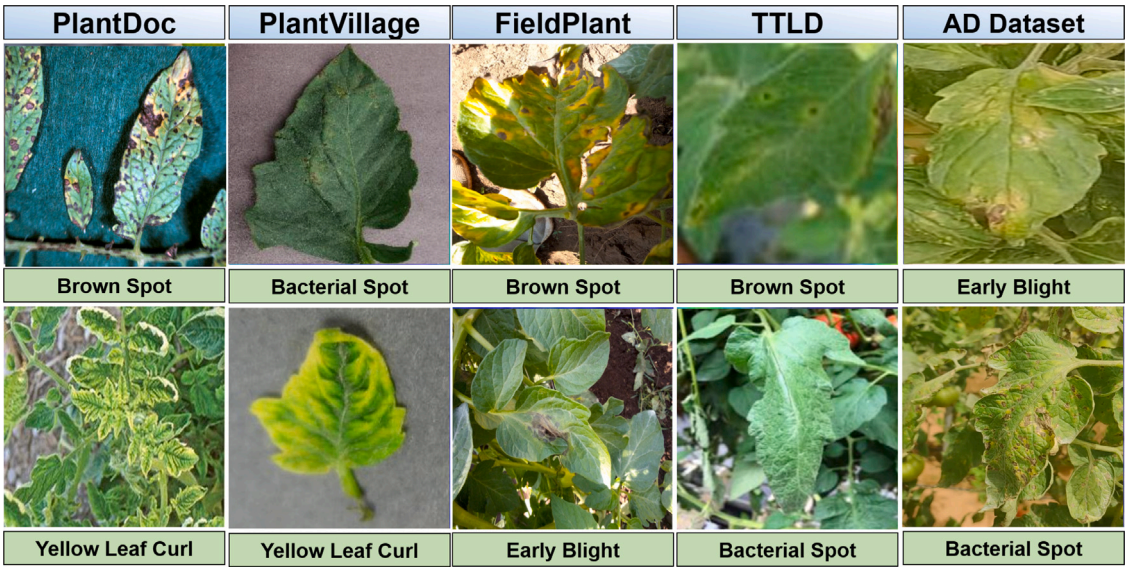

Fig. 4. Representative tomato leaf images from PlantDoc, PlantVillage, FieldPlant, TTLD, and the Abu Dhabi greenhouse dataset, illustrating disease symptoms across diverse imaging conditions.

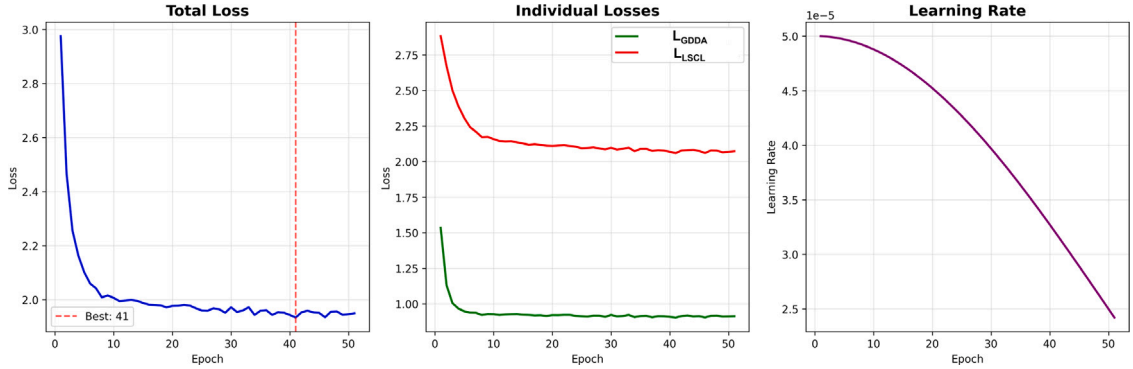

Fig. 5. Training dynamics of DLCAF. The total loss (left) shows stable convergence, while individual loss components (center) demonstrate balanced optimization of $\mathcal{L}_{\text{GDDA}}$ and $\mathcal{L}_{\text{LSCL}}$. The learning rate schedule (right) follows cosine decay.

optimizer with a learning rate of 1×10^{-5}, weight decay of 0.01, and cosine annealing scheduling. The model is trained for 50 epochs with a batch size of 8, employing early stopping with a patience of 10 epochs. Dropout (0.1) and label smoothing (0.1) are applied to improve generalization. Each training image is augmented using five fixed crops derived from expert-annotated lesion bounding boxes, supplemented with random crops when annotations are unavailable. Temperature parameters are set to $\tau_g = 0.05$ and $\tau_l = 0.07$, with the joint loss weight $\lambda = 0.5$. As shown in Fig. 5, DLCAF converges stably within 50 epochs, with balanced optimization of the GDDA and LSCL objectives. The total loss demonstrates smooth convergence, confirming the stability of joint cross-modal and lesion-aware training.

All experiments are implemented in PyTorch using the HuggingFace Transformers library and executed on an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM). Both vision and text encoders are initialized from pretrained CLIP weights, and the best model checkpoint is selected based on validation loss.

Evaluation is conducted under a zero-shot cross-dataset setting without fine-tuning on target domains. The vision encoder generates embeddings for each test image, which are compared against textual embeddings of disease descriptions using cosine similarity. The class with the highest similarity score is selected as the prediction. Performance is reported using overall accuracy and weighted F1-score to account for class imbalance, explicitly assessing the model's ability to generalize across datasets, including disease classes unseen during training.

5. Experimental results

5.1. Baseline comparison

Table 3 compares the proposed DLCAF framework with state-of-the-art vision–language models across four datasets with substantially different characteristics. PlantVillage consists of controlled laboratory images with uniform backgrounds, FieldPlant captures natural field conditions with variable illumination, diverse backgrounds, and occlusions, TTLD includes a mixture of laboratory- and field-acquired leaf images, and the AD Dataset comprises images collected under real greenhouse conditions in Abu Dhabi, UAE. All models are evaluated under a zero-shot setting using multi-template ensemble prompting.

DLCAF consistently outperforms all baseline models across all evaluated datasets. Among the baselines, LongCLIP achieves the highest accuracy and weighted F1-score on PlantVillage, EVA-CLIP performs best on TTLD, AgriCLIP yields the strongest baseline accuracy and weighted F1-score on FieldPlant, and SLIP achieves the highest baseline performance on the AD Dataset. Relative to these best-performing baselines, DLCAF improves accuracy by 12.38 percentage points on PlantVillage (+27.07%), 9.73 points on FieldPlant (+15.50%), 3.54 points on TTLD

Table 3
Zero-shot performance comparison with vision–language baselines using multi-template ensemble prompting (P1). Values in percentages with two decimal places. Best baseline is underlined, while DLCAF (Ours) shows overall best performance in **bold**.

Type	Model	PlantVillage		FieldPlant		TTLD		AD Dataset	
		Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
General	CLIP (Radford et al., 2021)	34.54	32.95	7.64	11.75	26.21	14.69	25.10	16.00
	SigLIP (Zhai et al., 2023)	12.25	11.50	15.37	22.21	13.67	5.65	16.73	11.55
	SLIP (Mu et al., 2021)	43.45	31.47	18.06	17.70	29.10	22.81	37.05	32.58
	EVA-CLIP (Sun et al., 2023)	21.07	18.79	18.52	22.59	32.64	27.31	22.75	18.47
	LongCLIP (Zhang et al., 2025)	45.73	35.20	15.86	21.54	31.99	23.39	21.10	14.10
Agri	AgriCLIP (Nawaz et al., 2025)	40.39	23.24	62.79	48.44	17.85	15.65	19.12	11.14
	SCOLD (Nguyen Quoc et al., 2026)	–	–	22.67	14.63	22.04	16.99	20.12	14.42
DLCAF (Ours)		**58.11**	**59.33**	**72.52**	**73.51**	**36.18**	**30.85**	**74.30**	**78.21**

(+10.85%), and 37.25 points on the AD Dataset (+100.54%). In terms of weighted F1-score, DLCAF achieves improvements of 24.13 points on PlantVillage (+68.56%), 25.07 points on FieldPlant (+51.75%), 3.54 points on TTLD (+12.96%), and 45.63 points on the AD Dataset (+140.08%).

The performance gains are most pronounced on FieldPlant and the AD Dataset, both of which represent challenging real-world conditions characterized by background clutter, illumination variability, and partial occlusion. These results indicate that coupling global image–text alignment with lesion-level symptom consistency is particularly effective under field and greenhouse conditions, where purely global alignment approaches often fail. Results for SCOLD on PlantVillage are omitted from Table 3 to ensure a fair comparison, as this dataset is included in the training regime of that model. Overall, the strong improvements observed on FieldPlant and the AD Dataset highlight DLCAF's ability to bridge the domain gap between controlled laboratory imagery and real agricultural environments. The results demonstrate that dual-level contrastive alignment, combined with lesion-aware supervision, enables robust generalization across laboratory, field, and greenhouse deployment scenarios.

5.2. Ablation studies

5.2.1. Prompt engineering strategies

Prompt formulation plays a critical role in zero-shot classification by influencing how textual descriptions align with visual features. We evaluate three prompting strategies to assess sensitivity to textual variation. Strategy P1 employs a multi-template ensemble with diverse formulations (e.g., "a close-up image of {class}", "plant pathology image of {disease characteristics}"), Strategy P2 uses generic templates such as "a photo of {class}", and Strategy P3 incorporates explicit disease-oriented descriptions (e.g., "a tomato leaf with {disease} showing characteristic symptoms").

Table 4 reports performance across prompting strategies, models, and datasets. Across the benchmark datasets, DLCAF achieves its strongest and most consistent performance under the multi-template ensemble strategy (P1), indicating that exposure to diverse textual descriptions during inference enhances zero-shot generalization. Replacing P1 with generic prompts (P2) leads to accuracy reductions of 3.73 points on PlantVillage, 20.03 points on FieldPlant, and 1.20 points on TTLD. A similar trend is observed when using disease-specific prompts (P3), with accuracy decreases of 2.87, 2.50, and 2.74 points on the same datasets, respectively. Despite these variations, DLCAF maintains relatively stable performance across all prompt formulations compared to baseline models.

On the AD Dataset, DLCAF also demonstrates strong robustness to prompt choice. While the multi-template ensemble (P1) yields the highest overall performance (74.30% accuracy), performance under P2 and P3 remains competitive, with accuracy reductions of 6.17 and 4.98 percentage points, respectively. This behavior contrasts with several baseline models, which exhibit substantial prompt sensitivity on the AD Dataset, particularly under generic or disease-specific prompts.

In contrast, baseline vision–language models exhibit pronounced sensitivity to prompt wording. For example, CLIP's accuracy on FieldPlant increases from 7.64% under P1 to 25.33% under P3, while SigLIP and LongCLIP display inconsistent behavior across datasets and prompting strategies. Results for SCOLD on PlantVillage are omitted from Table 4 to ensure a fair comparison, as this dataset is included in the training regime of that model. Overall, DLCAF demonstrates greater robustness to prompt variation than baseline models. This stability validates the proposed one-to-many semantic alignment strategy, enabling the model to learn semantically grounded representations that remain consistent across diverse textual formulations, in contrast to the fragile prompt dependence observed in single-caption vision–language models. To further quantify semantic robustness, we measured the standard deviation of accuracy across prompting strategies. DLCAF exhibits substantially lower variance compared to baseline models. For instance, on the AD dataset, DLCAF shows approximately 3% performance variance across prompts, whereas CLIP and LongCLIP fluctuate by more than 12%–14%. This reduced variance indicates improved semantic stability and decreased sensitivity to prompt formulation, mitigating a common source of inconsistency and overgeneralization in vision–language models.

5.2.2. Comparison with vision-centric models

Table 5 compares DLCAF with representative CNN- and transformer-based classifiers under a cross-dataset generalization setting, where all models are trained on PlantDoc and evaluated on shared disease classes across PlantVillage, FieldPlant, TTLD, and the AD Dataset.

As shown in Table 5, image-only models exhibit limited robustness under domain shift, with pronounced performance degradation under field and greenhouse conditions. ResNet-50 achieves the strongest baseline accuracy on FieldPlant, while ViT-Base performs best among vision-centric baselines on PlantVillage, TTLD, and the AD Dataset. However, these models show inconsistent behavior across environments and remain substantially outperformed by DLCAF.

In contrast, DLCAF consistently achieves the highest accuracy and weighted F1-score across all evaluated datasets. The performance gap is most pronounced on FieldPlant and the AD Dataset, where background clutter, illumination variability, and partial occlusions are prevalent. These results underscore the limitations of purely visual representations and highlight the benefit of integrating vision–language alignment with lesion-aware learning for robust cross-environment disease recognition.

5.2.3. Impact of expert annotations

A key distinction of DLCAF from standard contrastive learning approaches, such as SimCLR, is its use of expert-annotated bounding boxes to localize symptomatic lesions rather than relying on randomly sampled augmentations. This ablation evaluates whether professional lesion annotations provide more effective supervision for Local Symptom Consistency Learning (LSCL) than random crops extracted from full leaf images. Table 6 compares DLCAF trained with expert annotations against a variant using random crops. Expert annotations consistently outperform random cropping across all datasets. The largest gains

Table 4
Performance across prompting strategies. P1: multi-template ensemble, P2: generic templates, P3: disease-specific descriptions. Values in percentages with two decimal places.

Model	Strategy	PlantVillage		FieldPlant		TTLD		AD Dataset	
		Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
CLIP (Radford et al., 2021)	P1	34.54	32.95	7.64	11.75	26.21	14.69	25.10	16.00
	P2	32.50	29.63	13.70	20.83	25.56	12.84	57.37	42.48
	P3	14.57	17.64	25.33	13.00	15.27	12.70	26.69	18.59
SigLIP (Zhai et al., 2023)	P1	12.25	11.50	15.37	22.21	13.67	5.65	16.73	11.55
	P2	21.59	19.83	61.79	48.56	16.72	5.82	19.52	9.71
	P3	11.41	3.05	61.96	50.11	18.97	10.17	25.10	14.84
LongCLIP (Zhang et al., 2025)	P1	45.73	35.20	15.86	21.54	31.99	23.39	21.10	14.10
	P2	44.66	32.72	11.05	15.96	28.78	20.00	48.41	49.28
	P3	15.07	17.28	26.00	12.79	17.68	15.49	29.08	18.44
AgriCLIP (Nawaz et al., 2025)	P1	40.39	23.24	62.79	48.44	17.85	15.65	19.12	11.14
	P2	39.94	20.52	59.24	44.65	17.68	14.26	17.33	10.32
	P3	41.40	25.36	65.11	47.23	18.65	17.70	20.15	13.45
SCOLD (Nguyen Quoc et al., 2026)	P1	–	–	22.67	14.63	22.04	16.99	20.12	14.42
	P2	–	–	26.41	14.80	23.47	19.50	15.94	10.81
	P3	–	–	30.15	16.61	24.28	20.96	54.58	50.71
DLCAF (Ours)	P1	**58.11**	**59.33**	**72.52**	**73.51**	**36.18**	**30.85**	**74.30**	**78.21**
	P2	54.38	56.66	52.49	55.39	34.98	29.99	68.13	75.37
	P3	55.24	55.16	70.02	70.88	33.44	26.76	69.32	71.61

Table 5
Cross-dataset performance comparison across CNN- and transformer-based baseline models. Values are percentages with two decimal places. Best baseline is underlined, while DLCAF (Ours) shows overall best performance in **bold**.

Model	PlantVillage		FieldPlant		TTLD		AD Dataset	
	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
ConvNeXt-Tiny (Liu et al., 2022)	23.17	19.61	26.16	36.72	25.48	36.97	15.34	24.83
ResNet-50 (He et al., 2016)	14.93	4.99	54.65	49.78	17.83	19.17	0.80	1.40
EfficientNet-B0 (Tan and Le, 2019)	19.72	14.01	38.62	46.87	12.74	20.64	8.37	12.79
ViT-Base (Dosovitskiy et al., 2021)	38.31	38.74	6.40	10.23	43.31	55.20	45.02	58.41
Swin-Tiny (Liu et al., 2021)	37.85	39.34	11.88	18.46	26.75	39.06	26.10	37.54
DLCAF (Ours)	**65.89**	**69.50**	**72.52**	**73.51**	**63.06**	**60.82**	**74.30**	**78.21**

Table 6
Impact of crop selection strategy on LSCL performance. Expert annotations correspond to professionally identified symptomatic regions, while random crops sample arbitrary leaf areas. Values are reported in percentages with two decimal places.

Crop strategy	PlantVillage		FieldPlant		TTLD		AD Dataset	
	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
Expert Annotations	**58.11**	**59.33**	**72.52**	**73.51**	**36.18**	**30.85**	**74.30**	**78.21**
Random Crops	40.24	29.52	52.99	29.91	31.51	27.70	56.57	61.25

are observed on FieldPlant, where accuracy increases from 52.99% to 72.52% and weighted F1-score improves from 29.91 to 73.51. Similarly, on the AD Dataset, expert supervision yields a substantial performance improvement, with accuracy increasing from 56.57% to 74.30% and weighted F1-score from 61.25 to 78.21. These results validate the core design principle of DLCAF: explicit localization supervision provides more informative contrastive signals than arbitrary spatial sampling.

Random crops frequently include healthy tissue or background regions, leading to noisy positive pairs that weaken the class-level consistency enforced by LSCL. In contrast, expert annotations focus training on genuinely pathological regions, ensuring that crops from the same disease class capture consistent lesion characteristics. The magnitude of improvement varies across datasets. FieldPlant and the AD Dataset benefit most from expert supervision, reflecting environments where diseases manifest as localized lesions under visually complex field or greenhouse conditions. PlantVillage exhibits smaller gains, likely because certain diseases produce more global visual changes rather than sharply localized symptoms, while TTLD shows more modest improvements due to mixed imaging conditions and annotation uncertainty. Overall, these findings highlight the importance of lesion-aware supervision for robust, domain-adapted disease recognition in real-world agricultural deployments.

5.2.4. Effect of LSCL weight (λ)

Table 7 analyzes the impact of the relative weighting between the GDDA and LSCL objectives through the hyperparameter λ. The results illustrate the complementary roles of global semantic alignment and local symptom consistency, with dataset characteristics strongly influencing the optimal balance between the two. Setting $\lambda = 0$ disables local consistency learning, leaving only global disease-description alignment. This configuration produces strongly dataset-dependent behavior. FieldPlant maintains moderate performance (57.56% accuracy), whereas PlantVillage accuracy drops sharply to 22.09%. This contrast reflects differences in data characteristics. FieldPlant images contain complex backgrounds and illumination variations, where holistic leaf appearance provides informative contextual cues. In contrast, PlantVillage consists of isolated leaves against uniform backgrounds, making localized lesion patterns the primary discriminative signal. Without LSCL, GDDA alone fails to sufficiently emphasize these critical pathological regions.

Table 7
Impact of LSCL weight λ on zero-shot performance. Values are reported in percentages with two decimal places.

λ	PlantVillage		FieldPlant		TTLD		AD Dataset	
	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
0.0 (GDDA only)	22.09	18.79	57.56	62.55	31.58	25.06	69.72	71.25
0.5	58.11	59.33	**72.52**	**73.51**	**36.18**	30.85	74.30	78.21
1.0	60.79	60.67	63.49	66.01	35.22	**32.38**	56.77	62.84
2.0 (LSCL emphasis)	**64.75**	**63.87**	70.41	71.01	31.68	25.36	**75.10**	**79.18**

Table 8
Comparison between DLCAF and standard CLIP training. Values are reported in percentages with two decimal places.

Training method	PlantVillage		FieldPlant		TTLD		AD Dataset	
	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
CLIP (Radford et al., 2021)	33.20	31.14	56.23	64.22	31.99	28.71	68.21	67.34
DLCAF (Ours)	**58.11**	**59.33**	**72.52**	**73.51**	**36.18**	**30.85**	**74.30**	**78.21**

A balanced weighting of $\lambda = 0.5$ yields optimal or near-optimal performance across multiple datasets, substantially improving PlantVillage accuracy relative to the GDDA-only setting while preserving strong FieldPlant performance. This configuration also achieves strong performance on the AD Dataset, indicating that moderate lesion-level supervision provides an effective trade-off between local pathology modeling and global contextual cues. These results confirm that global semantic alignment and local symptom consistency interact synergistically, validating the dual-level optimization strategy of DLCAF. Increasing λ further reveals a non-linear performance trend. At $\lambda = 1.0$, accuracy improves on PlantVillage, reflecting stronger local feature discrimination. Pushing the weight to $\lambda = 2.0$ further benefits PlantVillage and yields the best performance on the AD Dataset, suggesting that greenhouse environments with relatively consistent backgrounds can profit from stronger lesion-level constraints. However, this trend does not generalize uniformly across all datasets.

On FieldPlant and TTLD, performance decreases at $\lambda = 2.0$, indicating that excessive emphasis on local crops can hinder generalization when global contextual cues remain informative. Under open-field conditions, both lesion morphology and broader scene context contribute to disease discrimination. Excessive lesion weighting can therefore disturb this balance, particularly when lesion appearance varies due to occlusions and viewpoint changes. For clarity, increasing λ strengthens the model's focus on lesion regions. Under $\lambda = 2.0$, this intensified lesion emphasis improves PlantVillage accuracy to 64.75%, as the controlled laboratory images contain clearly localized disease symptoms with minimal background interference. However, in cluttered field environments where contextual cues remain informative, excessive lesion weighting suppresses useful global signals, leading to performance degradation. Overall, these observations highlight the importance of balancing global and local objectives according to deployment environment characteristics.

5.2.5. Comparison with standard CLIP training

As shown in Table 8, DLCAF consistently outperforms standard CLIP-style training across all evaluated datasets. The improvements are most pronounced on PlantVillage and FieldPlant, where DLCAF achieves substantial gains in both accuracy and weighted F1-score, while also delivering consistent improvements on TTLD. On the AD Dataset, DLCAF further improves upon standard CLIP training, demonstrating the benefit of the proposed design under real greenhouse conditions.

These results demonstrate that extending CLIP-style training with one-to-many global semantic alignment and lesion-aware local supervision enables the model to learn more robust disease representations than conventional one-to-one image–text alignment alone. Overall, this ablation confirms that the performance gains achieved by DLCAF are not attributable to architectural differences alone, but stem from its dual-level contrastive learning strategy, which jointly leverages diverse disease descriptions and expert-guided symptom-level supervision.

5.2.6. Impact of training prompt diversity

Table 9 examines how the number of textual descriptions per disease class used during training affects zero-shot generalization. Across all datasets, the results exhibit a non-monotonic relationship between prompt diversity and performance. Training with five prompts per class consistently yields the best performance, while single-prompt training leads to notable reductions in both accuracy and weighted F1-score. Increasing the number of prompts beyond five does not provide additional gains and, in some cases, results in modest performance degradation.

A similar trend is observed on the AD Dataset, where performance improves as prompt diversity increases up to five descriptions per class and then saturates or slightly declines with further increases. This consistency across laboratory, field, and greenhouse datasets suggests that moderate prompt diversity provides sufficient semantic coverage for robust disease representation. This behavior reflects limitations associated with the frozen text encoder. While moderate prompt diversity enriches semantic supervision, increasing the number of prompts to seven or ten introduces diminishing returns, indicating a limited capacity of the pretrained encoder to distinguish fine-grained semantic variations. Beyond a certain threshold, additional prompts may introduce redundancy or semantic noise rather than meaningful diversity, destabilizing the learned representations. These results indicate the existence of an optimal level of semantic richness that balances linguistic variability with representational consistency when the text encoder is not adapted during training.

5.3. Uncertainty quantification and system reliability analysis

Vision–Language Models (VLMs) are known to exhibit overconfidence under domain shift (Xuan et al., 2025), which may compromise their reliability in safety-critical agricultural applications. To evaluate predictive calibration, we performed an uncertainty analysis using Expected Calibration Error (ECE), which measures the discrepancy between predicted confidence and empirical accuracy. Lower ECE values indicate better calibration and more reliable confidence estimates.

In addition to multi-class evaluation, we conducted a binary reliability analysis tailored to greenhouse deployment scenarios. All disease categories were grouped into a single "Diseased" class against a "Healthy" baseline. This formulation enables assessment of operational safety by quantifying the system's ability to detect infections while minimizing false negatives. In this context, the False Negative Rate (FNR) represents the proportion of infected samples incorrectly classified as healthy, which is critical for disease containment in controlled environments. Table 10 summarizes the calibration and reliability metrics of the proposed DLCAF framework across four datasets representing laboratory, field-based, and mixed acquisition conditions.

The results reveal a clear dependence between domain characteristics and model calibration. Since the VLM backbone was fine-tuned

Table 9
Impact of training prompt diversity on zero-shot performance. Values are reported in percentages with two decimal places.

Prompts/Class	PlantVillage		FieldPlant		TTLD		AD Dataset	
	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
1	45.31	47.95	52.99	59.30	25.73	20.55	45.22	52.31
3	49.44	50.68	47.52	54.71	27.01	24.69	55.78	61.99
5	**58.11**	**59.33**	**72.52**	**73.51**	**36.18**	**30.85**	**74.30**	**78.21**
7	46.79	47.76	62.54	68.85	33.13	28.35	61.16	68.24
10	51.91	50.85	71.88	72.36	34.41	30.15	59.16	60.61

Table 10
Uncertainty and reliability metrics of the proposed DLCAF framework evaluated across diverse laboratory, field, and mixed datasets.

Dataset	Binary accuracy	False Negative Rate (FNR)	ECE
PlantVillage	91.27%	5.97%	0.3790
FieldPlant	**97.09%**	**2.18%**	**0.1689**
TTLD	86.33%	1.74%	0.5597
AD Dataset	91.04%	0.00%	0.2299

primarily on PlantDoc, which consists of field-acquired images with natural illumination and complex backgrounds, the model demonstrates superior calibration in similar field-based environments. The FP dataset achieves the lowest ECE (0.1689), indicating well-aligned confidence estimates under conditions consistent with the training distribution.

In contrast, higher ECE values are observed in PlantVillage and TTLD, suggesting reduced calibration under domain shift. This behavior indicates that although classification accuracy remains competitive, the model's confidence estimates become less reliable when visual characteristics differ from those encountered during fine-tuning.

From a deployment perspective, the zero False Negative Rate observed on the AD dataset demonstrates strong infection detection capability in real greenhouse conditions. However, the increased calibration error in laboratory-style datasets underscores the importance of aligning training data distributions with target deployment environments to ensure both predictive accuracy and calibrated uncertainty. These reliability results also clarify concerns regarding the practical usability of TTLD performance. Although the multi-class accuracy under strict cross-domain transfer is lower (36.18%), the binary reliability metrics indicate stable infection detection behavior, with 86.33% Binary Accuracy and a low False Negative Rate (1.74%). For aggregation-based disease monitoring and management, reliable identification of infected regions is more critical than perfect subtype classification. From this operational perspective, the binary detection performance satisfies an acceptable reliability threshold for deployment-oriented surveillance. Furthermore, the inclusion of calibration analysis (ECE) directly addresses uncertainty and potential overgeneralization, providing a principled assessment of confidence alignment under domain shift.

5.4. Qualitative analysis

5.4.1. Attention map characterization

To qualitatively evaluate the interpretability and lesion-focused behavior of the proposed DLCAF framework, we generate Grad-CAM attention maps for representative samples and compare them with those produced by CLIP, ResNet, and ViT. Fig. 6 illustrates how each model allocates attention across symptomatic and non-symptomatic regions during prediction. Across all examples, DLCAF consistently concentrates attention on true pathological structures, including chlorotic patches, necrotic lesions, mosaic mottling, and ring-shaped spot patterns.

This behavior reflects the architectural design of DLCAF, in which Global Disease-Description Alignment (GDDA) aligns with disease-level semantics, while Local Symptom Consistency Learning (LSCL) enforces sensitivity to fine-grained lesion cues. Together, these objectives enable the model to jointly capture global contextual information and localized symptom morphology, resulting in attention maps that precisely highlight disease-relevant regions.

In contrast, baseline models frequently exhibit mislocalized attention. CLIP and ViT tend to distribute activation over large, non-diagnostic regions of the leaf or background, leading to incorrect predictions despite strong global representations. ResNet occasionally attends to high-contrast background textures rather than disease symptoms, indicating sensitivity to domain-specific visual noise. These misaligned activations correspond to the misclassification cases shown in Fig. 6. Overall, the comparison demonstrates that DLCAF not only improves predictive accuracy but also produces attention patterns that align more closely with expert-defined symptomatic regions, an important property for transparent and trustworthy agricultural decision-support systems.

5.4.2. System integration and field deployment

To evaluate the practical utility of the DLCAF-LLM framework, an integrated system for greenhouse-scale disease monitoring was developed. This prototype validates the end-to-end pipeline from initial data acquisition to final decision support. The complete field-level workflow and its integration with the analytical interface are illustrated in Fig. 7.

The experimental protocol follows a structured inspection cycle aligned with standard agronomic practices. Data acquisition involves stratified sampling of each tomato vine by capturing three images across the top, middle, and lower canopy levels. This approach ensures a comprehensive spatial assessment of the plant, accounting for the fact that disease symptoms are often non-uniformly distributed across the vertical strata.

The interactive web-based interface facilitates data management by allowing users to define the greenhouse layout with specific row counts and plant counts per row. To ensure the system is practical for large-scale deployment, we implemented a sequential acquisition protocol that utilizes chronological metadata. During inspection, the user captures images in a fixed sequence, following the greenhouse's physical row-by-row layout. Because modern smartphones assign unique, sequential timestamps to each file, the interface utilizes this metadata to automate the spatial mapping process. Upon upload, the system sorts the images by timestamp and assigns them to the predefined grid (e.g., the first three images are mapped to the first plant). For this pilot validation, we demonstrated the workflow using a representative 4×4 plant grid (48 images). While this setup serves as a controlled validation of the architectural logic, the system is fully scalable. The interface is agnostic to greenhouse dimensions, enabling rapid ingestion in larger facilities without requiring manual tagging or spatial-tracking hardware. Following automated ingestion, DLCAF processes each leaf independently to generate disease predictions. These results are aggregated at the plant level, where a plant is classified as infected if any of its associated leaf samples exhibit symptoms. As shown in Fig. 8, the interface visualizes the greenhouse health status as a spatial grid, enabling the rapid identification of localized disease hotspots.

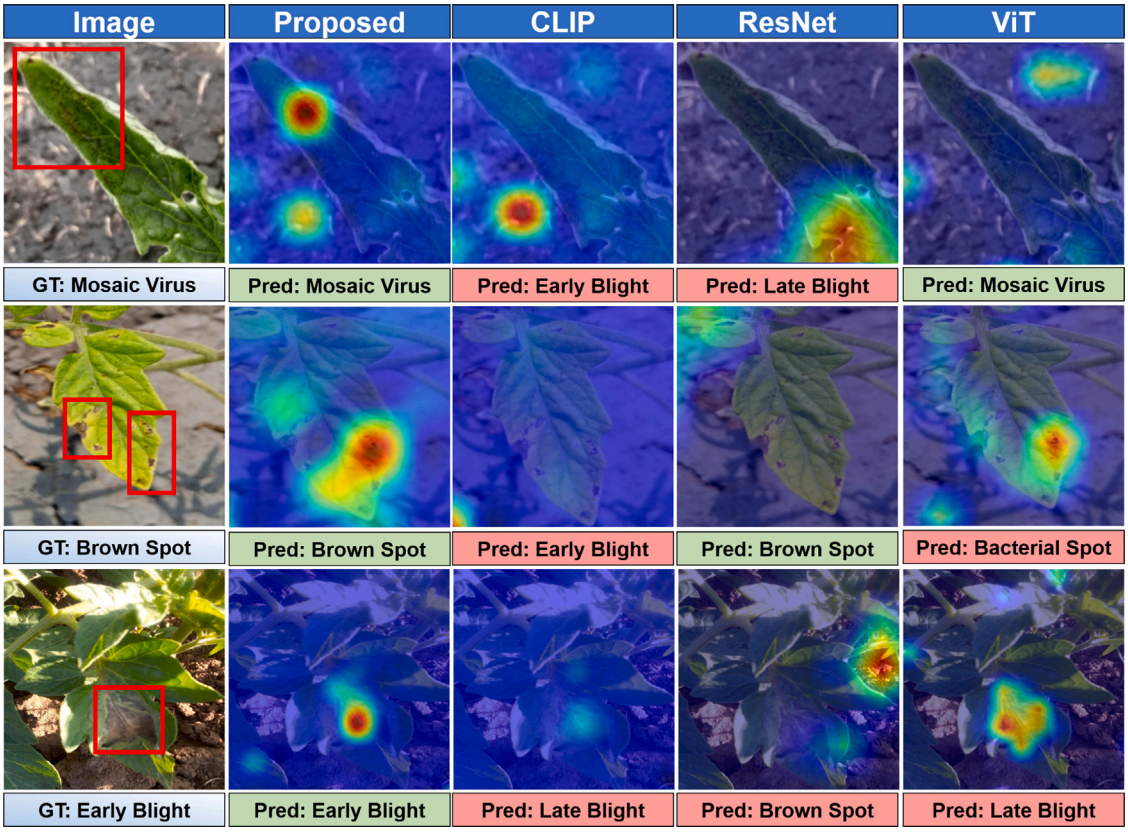

Fig. 6. Attention visualizations comparing the proposed DLCAF model with CLIP, ResNet, and ViT. For each sample, the first column shows the input image, followed by the corresponding attention maps. Ground-truth labels (GT) and predicted labels (Pred) are reported beneath each row. DLCAF consistently focuses on diagnostically relevant regions, whereas baseline models often highlight irrelevant areas or produce incorrect predictions.

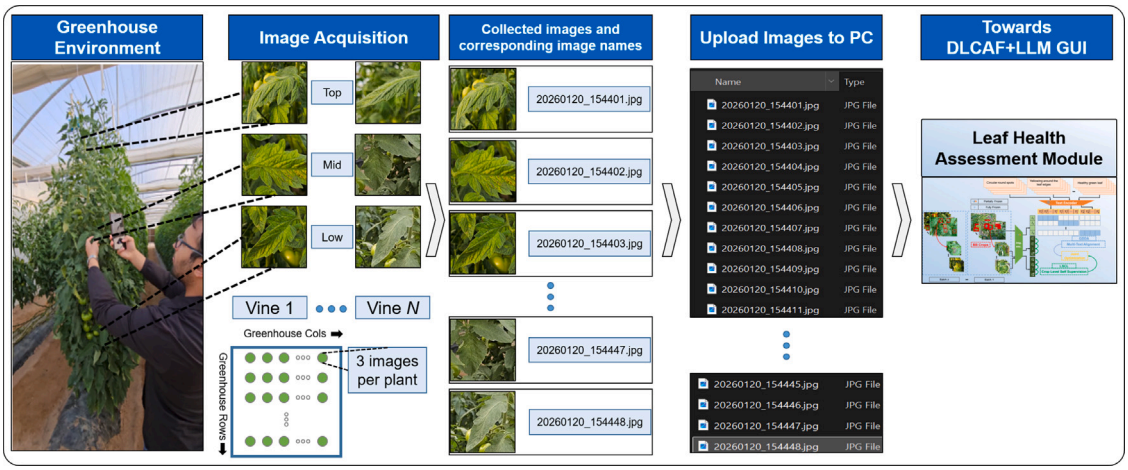

Fig. 7. Overview of the field experimental workflow and data ingestion process. The pipeline illustrates (left to right) the stratified image acquisition across vertical vine positions, the structured collection and naming conventions, and the subsequent data transfer to the Leaf Health Assessment module.

Beyond immediate visualization, the system computes and stores row-wise and greenhouse-level infection statistics for longitudinal analysis. When multiple observation cycles are recorded, an LLM-based reasoning module compares the stored summaries to identify temporal trends and assess disease progression. As illustrated in Fig. 9, the system generates an interpretable advisory that includes a risk assessment and a targeted action plan. This process transforms raw model predictions into decision-oriented guidance for effective greenhouse management.

6. Discussion

The superior performance of the DLCAF-LLM framework across diverse datasets underscores its ability to address two primary bottlenecks in automated plant pathology: linguistic ambiguity in disease characterization and spatial noise in field-level imagery. By bridging the gap between global semantic understanding and localized feature extraction, DLCAF offers a robust solution for real-world agricultural deployment.

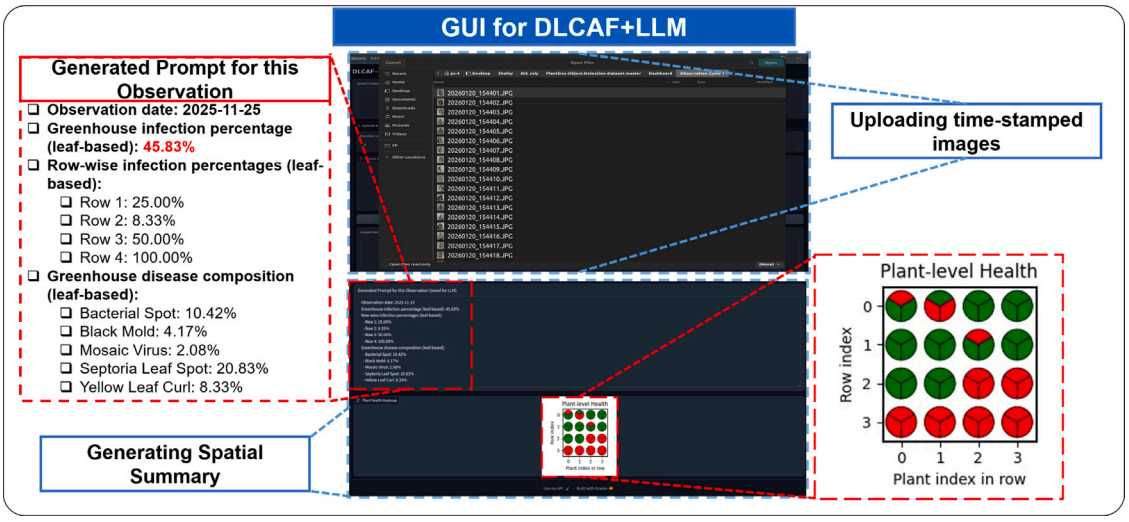

Fig. 8. Overview of the interactive GUI developed for the DLCAF+LLM framework. The interface supports structured image ingestion based on greenhouse layout, executes DLCAF for leaf-level disease recognition, aggregates predictions into plant- and row-level health maps, and stores spatial descriptors for subsequent temporal analysis.

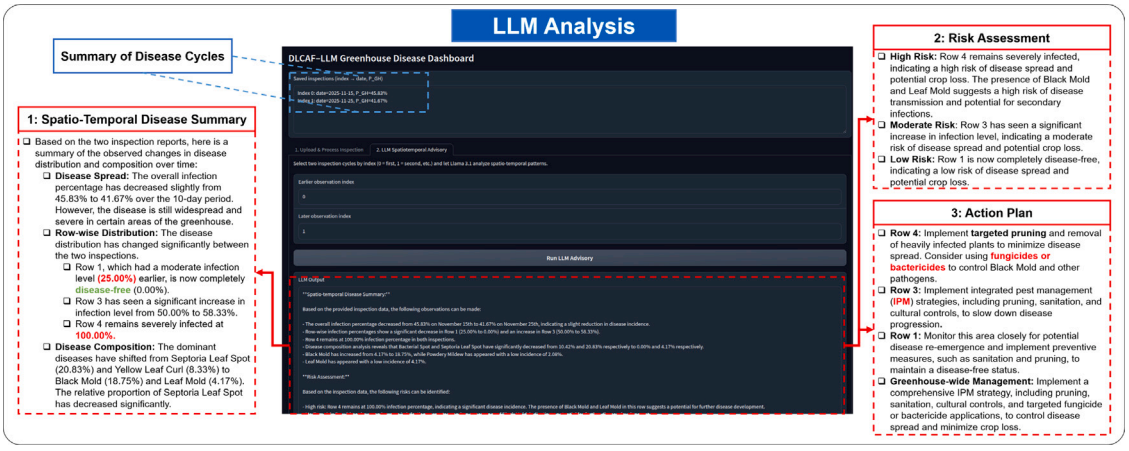

Fig. 9. Example of the LLM-driven spatiotemporal analysis module within the GUI. The system retrieves stored inspection summaries, compares selected inspection cycles, and generates an interpretable advisory comprising a spatiotemporal disease summary, risk assessment, and targeted action plan.

6.1. Impact of Dual-Level Contrastive Alignment

The core strength of DLCAF lies in the synergy between Global Disease Description Alignment (GDDA) and Local Symptom Consistency Learning (LSCL). Unlike conventional Vision–Language Models (VLMs) that rely on restricted one-to-one image–caption mappings (Radford et al., 2021; Zhai et al., 2023), GDDA employs a one-to-many alignment strategy. This allows the framework to capture the biological spectrum of disease, from early-stage chlorosis to advanced necrosis, within a flexible semantic embedding. By training on diverse expert descriptions, the model becomes resilient to linguistic variations, ensuring stable performance regardless of the specific phrasing used in diagnostic prompts.

Complementing this, LSCL enforces focus on localized pathological patterns. This mirrors the cognitive process of human experts who prioritize lesion morphology over holistic leaf appearance. Our results confirm that supervising specific disease-relevant regions provides a superior learning signal compared to the generic global supervision used in models like AgriCLIP (Nawaz et al., 2025) or SCOLD (Nguyen Quoc et al., 2026). By coupling these levels, DLCAF effectively decouples diagnostic features from irrelevant environmental artifacts and background noise.

6.2. Generalization across datasets and domain shift

The comparative results across FieldPlant, PlantVillage, AD, and TTLD reveal consistent patterns regarding domain complexity and representation bias. FieldPlant represents the most challenging setting due to illumination variability, background clutter, occlusions, and viewpoint changes. Under such open-field conditions, models that rely heavily on global appearance cues are more likely to capture spurious environmental correlations instead of disease-specific morphology. The substantial performance gap between DLCAF and baseline vision–language models on FieldPlant highlights the benefit of explicitly incorporating lesion-aware supervision under domain shift.

PlantVillage, in contrast, represents a controlled laboratory environment with isolated leaves and uniform backgrounds. The limited environmental variability reduces the likelihood of contextual overfitting, allowing generic models to achieve relatively strong performance. As a result, relative improvements appear smaller despite consistent gains. The AD dataset occupies an intermediate position, combining real-world imaging conditions with more structured backgrounds than open-field data. The strong performance of DLCAF in this setting further supports the advantage of integrating global semantic alignment with local symptom consistency.

To further clarify the observed performance differences, DLCAF achieves the highest accuracy on all evaluated datasets, including PlantVillage (58.11%), exceeding the strongest baseline (LongCLIP, 45.73%) by more than 12 percentage points. The variation in absolute accuracy across datasets therefore reflects inherent domain characteristics rather than instability of the proposed method. The comparatively larger gains in FieldPlant and AD indicate that emphasizing lesion-level representation through LSCL provides a robustness advantage under more complex environmental conditions, while maintaining competitive performance in controlled laboratory settings. Overall, these observations indicate that the proposed dual-level optimization enhances robustness as environmental complexity increases, while maintaining stable performance in controlled scenarios.

6.3. Comparison with existing vision–language models

While foundation models such as CLIP (Radford et al., 2021) and EVA-CLIP (Sun et al., 2023) provide powerful visual backbones, they lack the domain-specific granularity required for agronomy. Even agricultural adaptations like AgriCLIP and SCOLD, which incorporate domain-relevant text, remain limited by their reliance on whole-image supervision. These models are inherently sensitive to non-discriminative regions common in greenhouse settings. DLCAF addresses this gap by explicitly modeling lesion structure. By integrating LLM-based reasoning with these robust visual features, the framework transitions from simple classification to an interpretable diagnostic tool capable of generating actionable management recommendations.

6.4. Implications for greenhouse-level decision support

Beyond leaf-level recognition, the proposed framework addresses a critical gap in existing disease detection pipelines by enabling structured spatial aggregation and decision-oriented reasoning. By aggregating predictions into row- and greenhouse-level indicators, the system provides interpretable measures of infection pressure that are directly aligned with agricultural management practices. This spatial representation enables the identification of localized disease hotspots, comparison of inspection cycles, and monitoring of disease progression over time. The integration of a large language model further extends the utility of automated disease recognition by transforming quantitative indicators into actionable insights. Rather than producing isolated predictions, the LLM synthesizes spatial and temporal information to generate expert-like assessments, including risk evaluation, intervention prioritization, and monitoring recommendations. This capability bridges the gap between computer vision outputs and practical decision-making, a requirement that is often overlooked in purely classification-driven approaches. To mitigate the risk of hallucination or unsupported reasoning, the LLM component operates exclusively on structured numerical summaries derived from DLCAF predictions, such as disease prevalence, row-level statistics, and temporal comparisons. It does not interpret raw images nor independently generate diagnostic labels. This strict separation between perception (VLM-based recognition) and reasoning (LLM-based synthesis) constrains the LLM to verifiable quantitative inputs, thereby reducing the likelihood of speculative or uncontrolled inference in deployment settings.

6.5. Limitations and future directions

Despite its advantages, the proposed framework has several limitations. The reliance on expert-annotated lesion bounding boxes introduces scalability constraints, particularly for large-scale or resource-limited deployments. Although LSCL remains effective with randomly sampled crops, professional annotations consistently yield the strongest performance gains. Reducing dependence on expert supervision through weakly supervised or self-localizing strategies represents an important direction for future work.

Importantly, DLCAF is not structurally restricted to manual lesion annotations. The LSCL module operates on cropped regions and can integrate any region proposal mechanism. Automatic annotation pipelines based on foundation-model detectors (e.g., Grounding DINO Liu et al., 2025), open-world object detection frameworks, or segmentation models such as SAM can be employed to generate lesion or leaf proposals without manual intervention. These automatically extracted regions could replace expert bounding boxes in practical deployments, significantly reducing annotation cost while preserving the lesion-aware learning objective. Exploring automated region proposal strategies constitutes a natural extension for improving scalability and workflow efficiency.

Additionally, the use of a frozen text encoder constrains linguistic adaptability. While this design preserves pretrained semantic knowledge and stabilizes training, it may limit the model's ability to fully exploit evolving disease terminology or domain-specific language. Future research could explore partial or adaptive fine-tuning of language encoders to enhance flexibility without compromising robustness. Addressing these limitations will be essential for scaling vision–language decision-support systems across diverse crops, regions, and agricultural practices.

7. Conclusion

This work presented an end-to-end framework for tomato leaf disease management that integrates vision–language disease recognition, lesion-aware representation learning, spatial aggregation, and large language model (LLM)–based reasoning. At its core, the Dual-Level Contrastive Alignment Framework (DLCAF) jointly models global disease semantics and localized pathological patterns, enabling robust leaf-level disease recognition across diverse imaging conditions. By aggregating predictions across greenhouse rows and inspection cycles and interpreting them through an LLM, the proposed system extends automated disease detection into a practical, decision-oriented tool for greenhouse management. Extensive evaluation across laboratory, field, and mixed real-world datasets demonstrated that the proposed approach consistently outperforms existing vision-only and vision–language baselines, with particularly strong gains under challenging field conditions. These results highlight the importance of combining multi-description semantic alignment with lesion-focused learning to achieve robustness under domain shift. Beyond predictive accuracy, the framework provides interpretable spatial indicators and actionable management insights, addressing key requirements for real-world agricultural deployment. While the current implementation relies on expert-annotated lesion regions and a frozen language encoder, future work will explore scalable alternatives, including weakly supervised lesion localization and adaptive language modeling. Overall, this study establishes a principled foundation for integrating vision–language learning and large language models into decision-support systems for precision agriculture and sustainable greenhouse production.

CRediT authorship contribution statement

Muhammad Shafay: Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Divya Velayudhan:** Writing – review & editing, Methodology, Conceptualization. **Muhammad Owais:** Writing – review & editing, Validation, Conceptualization, Supervision. **Taimur Hassan:** Writing – review & editing, Visualization, Supervision. **Lakmal Seneviratne:** Supervision. **Irfan Hussain:** Writing – review & editing, Supervision, Project administration. **Naoufel Werghi:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used OpenAI's ChatGPT to assist in language refinement, formatting, and improving readability. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the final version of the manuscript.

Declaration of competing interest

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

Acknowledgments

This research was supported in part by the Khalifa University Center for Autonomous Robotic Systems (KUCARS) under Award No. RC1-2018-KUCARS. It was also supported in part by Silal Innovation Oasis under Grants. 8475000023, 8475000024, 8475000025, and 8475000026. The authors additionally acknowledge institutional support from the Department of Applied Artificial Intelligence and Robotics, Aston University, Birmingham, United Kingdom, and Abu Dhabi University (Ref: 19300904).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compag.2026.111692.

Data availability

The implementation code is publicly available at GitHub, along with our collected dataset released at Figshare.

References

Abouelmagd, L.M., Shams, M.Y., Marie, H.S., Hassanien, A.E., 2024. An optimized capsule neural networks for tomato leaf disease classification. EURASIP J. Image Video Process. 2024 (1), 2. http://dx.doi.org/10.1186/s13640-023-00618-9.

Chang, Y.-H., 2020. Dataset of Tomato Leaves. Mendeley, http://dx.doi.org/10.17632/NGDGG79RZB.1, [Online]. Available: https://data.mendeley.com/datasets/ngdgg79rzb/1.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.

Food and Agriculture Organization of the United Nations, 2023. Tomato Crop Information and Production Statistics. Tech. Rep., Food and Agriculture Organization of the United Nations (FAO), [Online]. Available: https://openknowledge.fao.org/bitstream/handle/20.500.12345/df90e6cf-4178-4361-97d4-5154a9213877.pdf. (Accessed: 20 October 2025).

George, R., Thuseethan, S., Ragel, R.G., Mahendrakumaran, K., Nimishan, S., Wimalasooriya, C., Alazab, M., 2025. Past, present and future of deep plant leaf disease recognition: A survey. Comput. Electron. Agric. 234, 110128. http://dx.doi.org/10.1016/j.compag.2025.110128, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169925002340.

Ghosh, D., Ghosh, S., Jana, N.D., Biswas, S., Mallipeddi, R., 2025. Designing optimal vision transformer architecture using differential evolution for tomato leaf disease classification. Comput. Electron. Agric. 238, 110824. http://dx.doi.org/10.1016/j.compag.2025.110824, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169925009305.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Huang, J., Hao, X., Wang, Y., Song, R., Mu, Z., Niu, S., Guo, X., 2026. Joint topic entity and intent recognition model for multimodal agricultural diseases and pests question answering. Comput. Electron. Agric. 241, 111253. http://dx.doi.org/10.1016/j.compag.2025.111253, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169925013596.

Kanwar, M., 2013. Performance of tomato under greenhouse and open field conditions in the trans-Himalayan region of India. Adv. Hortic. Sci. 25 (1), http://dx.doi.org/10.13128/AHS-12786, 2011. [Online]. Available: https://oaj.fupress.net/index.php/ahs/article/view/2941.

Karim, M.J., Goni, M.O.F., Nahiduzzaman, M., Ahsan, M., Haider, J., Kowalski, M., 2024. Enhancing agriculture through real-time grape leaf disease classification via an edge device with a lightweight CNN architecture and grad-CAM. Sci. Rep. 14 (1), 16022. http://dx.doi.org/10.1038/s41598-024-66989-9.

Khan, B., Das, S., Fahim, N.S., Banerjee, S., Khan, S., Al-Sadoon, M.K., Al-Otaibi, H.S., Islam, A.R.M.T., 2024. Bayesian optimized multimodal deep hybrid learning approach for tomato leaf disease classification. Sci. Rep. 14 (1), 21525. http://dx.doi.org/10.1038/s41598-024-72237-x.

Li, X., Yang, S., Li, S., 2026. From identification to prevention: a kiwi disease diagnosis system integrating vision and large language models. Comput. Electron. Agric. 244, 111484. http://dx.doi.org/10.1016/j.compag.2026.111484, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169926000797.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, IEEE Computer Society, Los Alamitos, CA, USA, pp. 9992–10002. http://dx.doi.org/10.1109/ICCV48922.2021.00986, [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. CoRR, abs/2201.03545. [Online]. Available: https://arxiv.org/abs/2201.03545.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L., 2025. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (Eds.), Computer Vision – ECCV 2024. Springer Nature Switzerland, Cham, pp. 38–55.

Mohanty, S.P., Hughes, D.P., Salathé, M., 2016. Using deep learning for image-based plant disease detection. Front. Plant Sci. Volume 7 - 2016, http://dx.doi.org/10.3389/fpls.2016.01419, [Online]. Available: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2016.01419.

Moupojou, E., Tagne, A., Retraint, F., Tadonkemwa, A., Wilfried, D., Tapamo, H., Nkenlifack, M., 2023. FieldPlant: A Dataset of Field Plant Images for Plant Disease Detection and Classification With Deep Learning. IEEE Access 11, 35398–35410. http://dx.doi.org/10.1109/ACCESS.2023.3263042.

Mu, N., Kirillov, A., Wagner, D., Xie, S., 2021. SLIP: Self-supervision meets language-image pre-training. http://dx.doi.org/10.48550/ARXIV.2112.12750, [Online]. Available: https://arxiv.org/abs/2112.12750.

Nawaz, U., Muhammad, A., Gani, H., Naseer, M., Khan, F.S., Khan, S., Anwer, R., 2025. AgriCLIP: Adapting CLIP for agriculture and livestock via domain-specialized cross-model alignment. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (Eds.), Proceedings of the 31st International Conference on Computational Linguistics. Association for Computational Linguistics, Abu Dhabi, UAE, pp. 9630–9639, [Online]. Available: https://aclanthology.org/2025.coling-main.644/.

Nguyen Quoc, K., Le Thi Thu, L., Quach, L.-D., 2026. A vision-language foundation model for leaf disease identification. Expert Syst. Appl. 299, 130084. http://dx.doi.org/10.1016/j.eswa.2025.130084, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417425037005.

Panno, S., Davino, S., Caruso, A.G., Bertacca, S., Crnogorac, A., Mandić, A., Noris, E., Matić, S., 2021. A review of the most common and economically important diseases that undermine the cultivation of tomato crop in the mediterranean basin. Agronomy 11 (11), http://dx.doi.org/10.3390/agronomy11112188, [Online]. Available: https://www.mdpi.com/2073-4395/11/11/2188.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning. ICML 2021, 18-24 July 2021, Virtual Event, In: Proceedings of Machine Learning Research, vol. 139, PMLR, pp. 8748–8763, [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html.

Satyanarayana, L.V., Rao, D.C., Nayak, S.K., 2025. Multi leaf disease identification and classification using efficient capsule convolutional shuffle attention network. Comput. Electron. Agric. 237, 110662. http://dx.doi.org/10.1016/j.compag.2025.110662, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169925007689.

Shafay, M., Hassan, T., Owais, M., Hussain, I., Khawaja, S.G., Seneviratne, L., Werghi, N., 2025. Recent advances in plant disease detection: challenges and opportunities. Plant Methods 21 (1), 140. http://dx.doi.org/10.1186/s13007-025-01450-0.

Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., Batra, N., 2020. PlantDoc: A dataset for visual plant disease detection. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. In: CoDS COMAD 2020, Association for Computing Machinery, New York, NY, USA, pp. 249–253. http://dx.doi.org/10.1145/3371158.3371196.

Song, Y., Li, M., Zhou, Z., Zhang, J., Du, X., Dong, M., Jiang, Q., Li, C., Hu, Y., Yu, Q., Wang, D., Dong, H., Yan, S., 2025. A lightweight method for apple disease segmentation using multimodal transformer and sensor fusion. Comput. Electron. Agric. 237, 110737. http://dx.doi.org/10.1016/j.compag.2025.110737, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169925008439.

Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y., 2023. EVA-CLIP: Improved training techniques for CLIP at scale. http://dx.doi.org/10.48550/ARXIV.2303.15389, [Online]. Available: https://arxiv.org/abs/2303.15389.

Sundhar, S., Sharma, R., Maheshwari, P., Kumar, S.R., Kumar, T.S., 2025. Enhancing leaf disease classification using GAT-GCN hybrid model. Front. Plant Sci. Volume 16 - 2025, http://dx.doi.org/10.3389/fpls.2025.1569821, [Online]. Available: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2025.1569821.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 97, PMLR, pp. 6105–6114, [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html.

Too, E.C., Yujian, L., Njuki, S., Yingchun, L., 2019. A comparative study of fine-tuning deep learning models for plant disease identification. Comput. Electron. Agric. 161, 272–279. http://dx.doi.org/10.1016/j.compag.2018.03.032, BigData and DSS in Agriculture, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169917313303.

Xiong, S., Wang, L., Zhang, Y., Dong, P., Wang, B., Che, Y., Shi, L., Si, H., 2025. Boosting crop disease recognition via automated image description generation and multimodal fusion. Comput. Electron. Agric. 239, 111082. http://dx.doi.org/10.1016/j.compag.2025.111082, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169925011883.

Xuan, W., Zeng, Q., Qi, H., Wang, J., Yokoya, N., 2025. Seeing is believing, but how much? A comprehensive analysis of verbalized calibration in vision-language models. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Suzhou, China, pp. 1408–1450. http://dx.doi.org/10.18653/v1/2025.emnlp-main.74, [Online]. Available: https://aclanthology.org/2025.emnlp-main.74/.

Yan, C., Liang, Z., Cheng, H., Li, S., Yang, G., Li, Z., Yin, L., Qu, J., Wang, J., Wu, G., Tian, Q., Yu, Q., Zhao, G., 2025. CDIP-ChatGLM3: A dual-model approach integrating computer vision and language modeling for crop disease identification and prescription. Comput. Electron. Agric. 236, 110442. http://dx.doi.org/10.1016/j.compag.2025.110442, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169925005484.

Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L., 2023. Sigmoid loss for language image pre-training. In: 2023 IEEE/CVF International Conference on Computer Vision. ICCV, IEEE Computer Society, Los Alamitos, CA, USA, pp. 11941–11952. http://dx.doi.org/10.1109/ICCV51070.2023.01100, [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01100.

Zhang, K., Ma, L., Cui, B., Li, X., Zhang, B., Xie, N., 2024. Visual large language model for wheat disease diagnosis in the wild. Comput. Electron. Agric. 227, 109587. http://dx.doi.org/10.1016/j.compag.2024.109587, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169924009785.

Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J., 2025. Long-clip: unlocking the long-text capability of clip. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (Eds.), Computer Vision – ECCV 2024. Springer Nature Switzerland, Cham, pp. 310–325.

Zhu, H., Shi, W., Guo, X., Lyu, S., Yang, R., Han, Z., 2025. Potato disease detection and prevention using multimodal AI and large language model. Comput. Electron. Agric. 229, 109824. http://dx.doi.org/10.1016/j.compag.2024.109824, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169924012158.