Progressive Layer Activation CLIP for Few-Shot and Generalizable Cassava Disease Recognition

Muhammad Shafay

100057573@ku.ac.ae

Khalifa University

Muhammad Owais

Khalifa University

Divya Velayudhan

Khalifa University

Taimur Hassan

Abu Dhabi University

Irfan Hussain

Khalifa University

Naoufel Werghi

Khalifa University

Research Article

Keywords: Few-shot learning, Progressive layer activation, CLIP adaptation, Attention visualization

Posted Date: March 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9109616/v1

Additional Declarations: The authors declare no competing interests.

Progressive Layer Activation CLIP for Few-Shot and Generalizable Cassava Disease Recognition

Muhammad Shafay*, Muhammad Owais†, Divya Velayudhan*, Taimur Hassan‡, Irfan Hussain†, Naoufel Werghi*
*Department of Computer Science, Khalifa University, Abu Dhabi, UAE
Email: {100057573, divya.velayudhan, naoufel.werghi}@ku.ac.ae
†Khalifa University Center for Autonomous Robotic Systems, Khalifa University, Abu Dhabi, UAE
Email: {muhammad.owais, irfan.hussain}@ku.ac.ae
‡Department of Electrical & Computer Engineering, Abu Dhabi University, Abu Dhabi, UAE
Email: taimur.hassan@adu.ac.ae

***Abstract*—Cassava diseases such as Cassava Mosaic Disease (CMD), Cassava Brown Streak Disease (CBSD), and Cassava Bacterial Blight (CBB) pose serious threats to global food security, particularly in resource-limited regions where expert diagnosis is scarce. Although large vision–language models enable automated plant disease recognition, existing fine-tuning approaches struggle under extreme data scarcity. This paper proposes Progressive Layer Activation CLIP (PLA-CLIP), a curriculum-inspired fine-tuning framework for efficient few-shot classification of cassava diseases. PLA-CLIP progressively unfreezes transformer layers during training, stabilizing the optimization process while preserving pretrained vision–language alignment. Using only 43 images per class, PLA-CLIP achieves 78.25% accuracy and a 78.00% F1-weighted score on CD1, outperforming zero-shot CLIP by +15.98% and standard fine-tuning by +3.94%. Cross-dataset evaluations on CD2 and CD3 demonstrate robust generalization across varying conditions. Attention map visualizations confirm that the model focuses on disease-relevant regions, supporting interpretability. With a 2.65 ms inference time and moderate model size, PLA-CLIP offers an effective balance between efficiency and performance for practical plant health monitoring. The implementation and experimental code are publicly available at https://github.com/mshafay5/PLA-CLIP.**

***Index Terms*—Few-shot learning, Progressive layer activation, CLIP adaptation, Attention visualization**

I. Introduction

Cassava (*Manihot esculenta* Crantz) is a staple crop and the third-largest global source of carbohydrates, supporting over 500 million people in developing regions [1]. Its production is severely affected by diseases such as Cassava Mosaic Disease (CMD), Cassava Brown Streak Disease (CBSD), Cassava Bacterial Blight (CBB), and Cassava Green Mottle (CGM), causing annual yield losses exceeding US$1 billion [2]. Disease diagnosis is traditionally performed through expert visual inspection, which is slow, subjective, and difficult to scale in resource-limited areas [3].

Although deep learning has advanced automated plant disease recognition, most existing methods depend on large labeled datasets and substantial computational resources, limiting their applicability in crop-specific and low-resource settings [4], [5]. Similar challenges—such as data imbalance, occlusion, and limited annotations—have been observed in other constrained vision domains, motivating the development of structured and progressive learning strategies to improve robustness under data scarcity [6]–[8]. However, these approaches typically rely on task-specific, unimodal architectures and fully supervised training.

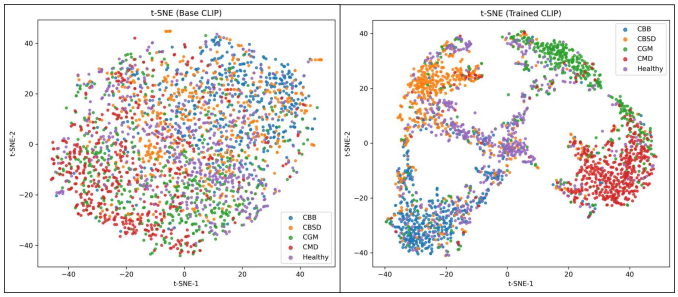

Fig. 1. t-SNE visualization of feature embeddings for the CD1 test set. Compared to pretrained CLIP, PLA-CLIP produces more compact and well-separated clusters after progressive adaptation, indicating improved discriminative feature learning under few-shot conditions.

Vision–language models offer strong zero-shot and few-shot capabilities through large-scale contrastive pretraining [9], [10], yet standard fine-tuning and adapter-based methods often underperform when training data are extremely limited. Recent studies indicate that structured adaptation and staged learning within large multimodal models can improve robustness and generalization under limited supervision [11], motivating the exploration of curriculum-driven adaptation strategies for vision–language models in specialized domains.

To address these challenges, this paper proposes Progressive Layer Activation CLIP (PLA-CLIP), a curriculum-inspired fine-tuning framework for few-shot classification of plant diseases. PLA-CLIP progressively unfreezes transformer layers during training, stabilizing the optimization process while preserving the pretrained vision–language alignment. Using only 43 images per class, PLA-CLIP achieves 78.25% accuracy and a 78.00% F1-weighted score on CD1, outperforming zero-shot CLIP by +15.98% and conventional fine-tuning by +3.94%. Cross-dataset evaluations on CD2 and CD3 demonstrate strong generalization across diverse conditions. Additionally, t-SNE visualizations (Figure 1) show compact and well-separated feature clusters, confirming the effectiveness and data efficiency

of the representation learning.

A. Related Work

Early plant disease classification relied on domain-specific convolutional neural networks (CNNs) trained on large labeled datasets. Ramcharan et al. [1] demonstrated strong performance for cassava disease detection using transfer learning, but at the cost of extensive data and computation. Subsequent approaches have improved efficiency [12]–[14], yet they remain data-intensive and unsuitable for edge deployment.

The scarcity of diverse, high-quality annotated data remains a major bottleneck for real-world agricultural vision systems. Recent dataset efforts emphasize realistic field conditions, environmental variability, and operational constraints [15], highlighting the limitations of purely supervised learning. A comprehensive review [16] reports a significant performance gap between laboratory benchmarks and real-field deployment, underscoring the importance of cross-dataset evaluation and robust generalization. This study directly addresses this issue by performing evaluations on three distinct datasets.

Vision–language models, such as CLIP [9], enable scalable zero-shot learning via image–text contrastive pretraining. Several adaptation methods, including CoOp [17], CLIP-Adapter [18], and Tip-Adapter [19], extend CLIP to few-shot scenarios but often struggle in specialized domains with limited data. Domain-aligned strategies [20] improve transferability at the cost of increased complexity.

Curriculum-based fine-tuning has recently shown promise for stabilizing adaptation in vision transformers [21]. Building on this principle, PLA-CLIP introduces progressive layer activation for vision–language models, enabling controlled capacity expansion during adaptation and achieving consistent accuracy gains (+2–4%) over standard fine-tuning while maintaining efficiency and strong transferability for agricultural visual recognition.

II. Methodology

A. Problem Formulation

We address few-shot cassava disease classification under extreme data scarcity using Progressive Layer Activation CLIP (PLA-CLIP). As shown in Figure 2, only 1% of the dataset (43 images per class, 215 total) is used for training, while the remaining 99% (≈21K images) is reserved for testing. The objective is to adapt CLIP's pretrained vision–language representations to the agricultural domain through progressive layer activation, enabling effective learning with minimal supervision.

During training, images and corresponding disease descriptions are encoded using CLIP image and text encoders to produce embeddings $\{I_i\}_{i=1}^{n}$ and $\{T_j\}_{j=1}^{m}$, where $m = 5$ denotes the disease classes (CBB, CBSD, CGM, CMD, and Healthy). Contrastive learning aligns matched image–text pairs while separating mismatched pairs, enforcing a shared semantic space. During inference, cosine similarity between image embeddings and class-specific text prompts determines the predicted class. Attention map visualizations further confirm that the model focuses on disease-affected leaf regions, supporting interpretability.

B. Progressive Layer Activation Strategy

PLA-CLIP employs a curriculum-based progressive layer activation strategy, gradually increasing model capacity during training instead of fine-tuning all parameters simultaneously. This approach stabilizes optimization, mitigates overfitting, and enables controlled adaptation under few-shot constraints. Transformer layers in the vision encoder are incrementally unfrozen according to the following schedule:

$$\text{ActiveLayers}(s) = \begin{cases} \{T_9 - T_{11}\}, & s = 1, \\ \{T_6 - T_{11}\}, & s = 2, \\ \{T_3 - T_{11}\}, & s = 3, \\ \{T_0 - T_{11}\}, & s = 4, \end{cases}$$

where s denotes the training stage and T_k represents the k-th transformer block.

This design leverages the hierarchical structure of Vision Transformers: early layers capture generic visual features, while deeper layers encode higher-level semantics. By initially fine-tuning only upper layers, PLA-CLIP preserves robust low-level representations and avoids catastrophic forgetting. Gradual activation of deeper layers enables fine-grained domain adaptation once higher-level features stabilize, achieving a balance between specialization and training stability under extreme data scarcity.

C. CLIP Contrastive Learning

To preserve vision–language alignment during adaptation, PLA-CLIP adopts the original CLIP contrastive objective. Given paired image–text samples, the loss maximizes similarity between matching pairs while minimizing similarity between mismatched pairs. For normalized embeddings of image I_i and text T_i, the symmetric contrastive loss is defined as:

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(I_i, T_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(I_i, T_j)/\tau)}, \tag{1}$$

$$\mathcal{L}_{\text{t2i}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(T_i, I_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(T_i, I_j)/\tau)}, \tag{2}$$

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2}\left(\mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}\right), \tag{3}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a learnable temperature, and N is the batch size.

The bidirectional formulation maintains multimodal consistency throughout progressive unfreezing, enabling effective specialization for plant disease recognition while preserving CLIP's pretrained semantic grounding and transferability across datasets.

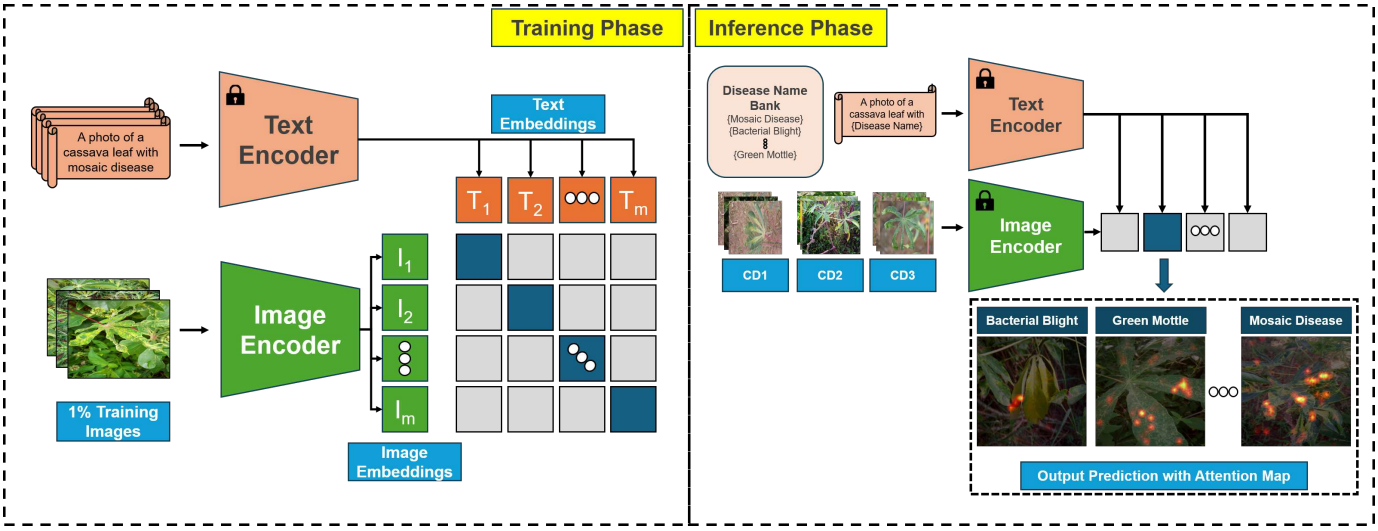

Fig. 2. Overview of the proposed PLA-CLIP framework. During training (left), image–text embeddings are aligned using contrastive learning while progressively unfreezing transformer layers across four stages. During inference (right), disease classes are predicted via image–text similarity, and attention maps highlight infected regions.

D. Training Configuration

PLA-CLIP is trained for 50 epochs using the AdamW optimizer with a learning rate of 1×10^{-5} for CLIP parameters. Progressive layer unfreezing occurs at epochs 10 (Stage 2), 20 (Stage 3), and 30 (Stage 4). A multi-step learning rate scheduler reduces the rate by a factor of 0.5 at epochs 15, 30, and 40, synchronized with the progressive unfreezing schedule to ensure smooth convergence. The batch size is set to 16, and the contrastive temperature parameter is initialized to $\tau = 0.07$.

Throughout training, the text encoder remains completely frozen to preserve its linguistic representations, while only the vision encoder is progressively adapted to the target domain. This asymmetric training strategy reflects the observation that textual disease descriptions require minimal adaptation, whereas visual representations demand greater specialization for plant pathology tasks. Freezing the text branch prevents linguistic knowledge degradation, allowing the model to focus on refining visual features. This design choice also enhances computational efficiency, making PLA-CLIP suitable for deployment in automated, resource-constrained agricultural systems.

E. Experimental Setup

We evaluated PLA-CLIP on three cassava leaf disease datasets, denoted CD1 [22], CD2 [23], and CD3 [24], each representing different environments and imaging conditions. CD1, the primary dataset, comprises five classes (CBB, CBSD, CGM, CMD, and Healthy) collected under field conditions and is used for training and testing purposes. Only 1% of its data (43 images per class, 215 total) is used for training, while the remaining 99% (21,182 images) serve as the test set.

The CD2 (Makerere University Cassava Dataset) comprises over 9,000 field images from Uganda, each with bounding box annotations for CBSD, CMD, and Healthy leaves, reflecting realistic agricultural conditions. CD3 (Cassava Leaf Disease Dataset) comprises RGB images (512×512) of CBB, CMD, and Healthy leaves captured in Tamil Nadu, India, under natural lighting. CD2 and CD3 are used exclusively for the evaluation of zero-shot cross data to test generalization without fine-tuning. Performance is reported using Accuracy and Weighted F1-score.

TABLE I
CASSAVA DATASETS USED FOR PLA-CLIP EVALUATION.

Dataset	Classes	Usage
CD1	CBB, CBSD, CGM, CMD, Healthy	Train (1%) + Test (99%)
CD2	CBSD, CMD, Healthy	Cross-test
CD3	CBB, CMD, Healthy	Cross-test

All experiments use the CLIP ViT-B/16 backbone with the AdamW optimizer, batch size 16, and a learning rate of 1×10^{-5}. Training follows the progressive layer activation schedule for 50 epochs. The images are resized to 224×224 and normalized using CLIP preprocessing. Comparisons include Zero-shot CLIP [9], CoOp [17], CLIP-Adapter [18], FT-CLIP, ResNet50 [25], ViT-Base [26], Swin Transformer [27], MobileNet [28], and EfficientNet [29]. All models share identical settings, and results are averaged over three random seeds for consistency.

III. RESULTS

A. Training Duration Analysis

To establish a baseline for the adaptation of a few-shots, the standard fine-tuned CLIP (FT-CLIP) model was trained in only 1% of CD1 (43 images per class) for two different durations: 10 and 50 epochs. As shown in Table II, extending the training from 10 to 50 epochs produces a small but consistent improvement. Accuracy increases from 73.59% to

74.31%, and the F1-weighted score also improves, indicating that longer optimization helps stabilize convergence and refine the learned representations. Compared with the zero-shot CLIP baseline (62.27%), both configurations demonstrate substantial performance gains under extreme data scarcity. These results form the reference point against which PLA-CLIP is evaluated, highlighting the additional benefits introduced by progressive layer activation.

TABLE II
PERFORMANCE OF FINE-TUNED CLIP TRAINED ON 1% OF CD1 WITH DIFFERENT TRAINING DURATIONS. LONGER TRAINING LEADS TO IMPROVED CONVERGENCE AND FEATURE REFINEMENT.

Configuration	Accuracy (%)	F1-weighted (%)
Zero-shot CLIP	62.27	48.83
FT-CLIP (10 epochs)	73.59	75.06
FT-CLIP (50 epochs)	74.31	75.75

B. Comprehensive Method Comparison

Table III summarizes the performance of PLA-CLIP compared with both vision–language and conventional computer vision baselines. All models were trained using only 1% of the CD1 dataset (43 images per class, totaling 215 images) and evaluated on the remaining 99% (21,182 images), simulating an extreme few-shot learning scenario. Among vision–language approaches, direct fine-tuning or naïve adaptation tends to reduce generalization. For example, CoOp (58.60%) and CLIP-Adapter (59.40%) perform below the zero-shot CLIP baseline (61.27%), suggesting that unstructured parameter updates can disrupt the pretrained alignment between visual and textual representations.

The proposed PLA-CLIP achieves 78.25% Accuracy and 78.00% F1-weighted score, surpassing both standard fine-tuned CLIP (74.31%) and the strongest traditional baseline, ViT-Base (73.96%). These results confirm that progressive layer activation enables stable and data-efficient adaptation under limited supervision. By gradually expanding model capacity, PLA-CLIP retains general visual knowledge while enhancing domain-specific representation learning, establishing a new benchmark for few-shot cassava disease classification.

C. Schedule Sensitivity Analysis

To evaluate whether the progressive activation mechanism depends on a particular staging configuration, a schedule sensitivity study was conducted on CD1 using only 1% of the data for training and the remaining 99% for evaluation. Three progressive unfreezing variants, 3-stage, 4-stage, and 5-stage, were trained for 50 epochs under identical settings. As shown in Table IV, the 4-stage configuration, which corresponds to the main PLA-CLIP design, achieved the highest accuracy (78.25%) and F1-weighted score (78.00%), outperforming both the 3-stage and 5-stage schedules.

These findings show that the improvements achieved by PLA-CLIP are driven by the progressive activation strategy itself rather than by a particular hyperparameter choice. The comparable performance of the 3-stage and 4-stage schedules highlights the robustness of the approach, whereas the decline observed with the 5-stage variant suggests that excessively granular unfreezing may introduce instability and overfitting under severe data limitations. Overall, the 4-stage configuration offers the best balance between adaptability and stability for a few-shot fine-tuning on CD1.

TABLE III
PERFORMANCE COMPARISON OF PLA-CLIP WITH EXISTING BASELINES. ALL MODELS WERE TRAINED ON 1% OF CD1 AND EVALUATED ON THE REMAINING 99%. CLIP-BASED METHODS USE THE VIT-B/16 BACKBONE.

Method	Accuracy (%)	F1-weighted (%)
Vision–Language Approaches		
Zero-shot CLIP [9]	61.27	48.83
CoOp [17]	58.60	62.13
CLIP-Adapter [18]	59.40	56.47
FT-CLIP	74.31	75.75
Traditional Computer Vision		
ResNet50 [25]	65.40	67.03
ViT-Base [26]	73.96	74.16
Swin Transformer [27]	72.28	73.46
MobileNet [28]	51.43	55.85
EfficientNet [29]	43.97	47.84
Proposed Method		
PLA-CLIP (Ours)	**78.25**	**78.00**

TABLE IV
SCHEDULE SENSITIVITY ANALYSIS ON CD1 (50 EPOCHS), TRAINED ON 1% AND EVALUATED ON 99%.

Schedule	Accuracy (%)	F1-weighted (%)
3 stage	72.93	74.84
4 stage	**78.25**	**78.00**
5 stage	67.29	70.39

D. Cross-Dataset Generalization

To assess the robustness of PLA-CLIP beyond the training domain, the model trained on CD1 was evaluated directly on two unseen datasets, CD2 and CD3, without additional fine-tuning. These datasets originate from different geographic regions and exhibit distinct environmental conditions, providing a realistic measure of domain generalization. As shown in Table V, PLA-CLIP achieved strong performance on CD2 with an accuracy of 75.85% and a weighted F1-score of 81.50%, indicating effective transfer of visual–textual representations to new data distributions.

TABLE V
CROSS-DATASET GENERALIZATION COMPARISON BETWEEN THE BASE CLIP MODEL AND THE PROPOSED PLA-CLIP, TRAINED ON CD1 AND EVALUATED ON UNSEEN DATASETS CD2 AND CD3 WITHOUT ADDITIONAL FINE-TUNING.

Dataset	**Model**	**Accuracy (%)**	**F1-weighted (%)**
CD2	Zero-shot CLIP	54.09	38.59
CD2	PLA-CLIP (Ours)	**75.85**	**81.50**
CD3	Zero-shot	37.28	21.69
CD3	PLA-CLIP (Ours)	**52.63**	**62.11**

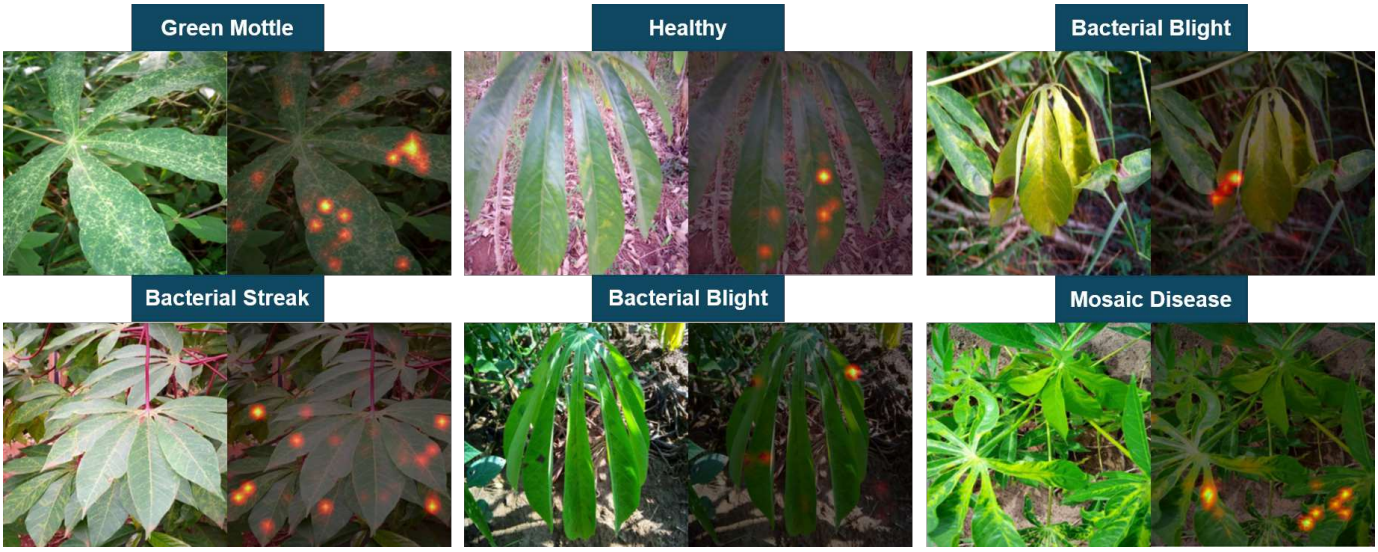

Fig. 3. Attention map visualization for different cassava leaf disease categories. The highlighted regions correspond to areas of model focus during prediction, demonstrating that PLA-CLIP accurately attends to disease-affected regions while maintaining low activation on healthy samples.

In CD3, performance was comparatively lower (52.63% accuracy and 62.11% F1-score), primarily due to the smaller sample size and the significant domain gap between Indian and Ugandan field conditions. However, PLA-CLIP consistently outperformed the base CLIP model (54.09% on CD2 and 37.28% on CD3), highlighting that the progressive layer activation strategy improves representation transfer, stability, and robustness under few-shot constraints.

E. Attention Map Visualization

To better interpret the model's predictions and assess whether it focuses on relevant regions of the image, we visualize the attention maps generated by PLA-CLIP for different classes of cassava disease. As shown in Figure 3, the highlighted regions correspond closely to the diseased areas of the leaves, indicating that the model effectively captures discriminative visual cues. For example, in cases of Bacterial Blight and Cassava Mosaic Disease, the attention heatmaps align with visible lesions and chlorotic patterns. In contrast, for healthy samples, the focus remains distributed over the leaf surface without concentrated activations. These visualizations confirm that PLA-CLIP not only improves classification accuracy under few-shot conditions but also enhances interpretability by learning semantically meaningful and disease-specific representations.

F. Model Efficiency and Practical Considerations

Table VI presents a comparison of the inference efficiency of PLA-CLIP with commonly used computer vision architectures. PLA-CLIP achieves a balanced trade-off between computational cost and performance, with an inference time of 2.65 ms and a model size of 570.91 MB. Although larger than lightweight networks such as MobileNetV2 and EfficientNet-B0, PLA-CLIP delivers substantially higher accuracy and feature robustness, demonstrating that its computational overhead is justified when precision is critical. Compared with transformer-based baselines, it operates faster than the Swin Transformer while maintaining similar FLOPs and achieving superior accuracy. These results highlight that PLA-CLIP provides an efficient compromise between scalability and accuracy, making it suitable for research and applied settings where reliable and interpretable predictions are prioritized over minimal computational cost.

TABLE VI
INFERENCE EFFICIENCY COMPARISON OF PLA-CLIP WITH STANDARD COMPUTER VISION MODELS.

Model	Params (M)	GFLOPs	Inference (ms)	Size (MB)
ResNet50	25.56	4.11	1.41	97.80
ViT-Base	86.57	16.87	2.37	330.29
Swin-Base	87.77	15.47	5.36	334.94
MobileNetV2	3.50	0.31	1.29	13.61
EfficientNet-B0	5.29	0.40	1.99	20.46
PLA-CLIP (Ours)	149.62	16.87	2.65	570.91

IV. CONCLUSION

This study presented Progressive Layer Activation CLIP (PLA-CLIP), a data-efficient vision–language framework designed for the classification of plant diseases in a few-shots under extreme data scarcity. The proposed approach leverages a curriculum-inspired progressive layer activation strategy that incrementally unfreezes transformer layers, allowing the model to adapt smoothly to new domains while maintaining the pretrained multimodal alignment of CLIP. Through this gradual capacity expansion, PLA-CLIP achieved stable convergence, enhanced feature specialization, and improved cross-domain generalization. Comprehensive experiments demonstrated that PLA-CLIP consistently outperforms both standard fine-tuning and training-free adaptation methods. The model achieved 78.25% accuracy and 78.00% F1-weighted score

on CD1, representing a +15.98% improvement over zero-shot CLIP and a +3.94% gain over traditional fine-tuning. Cross-data evaluations on CD2 and CD3 further validated the robustness and ability of the model to generalize effectively to unseen geographic and environmental conditions. Interpretability analysis using attention map visualizations revealed that PLA-CLIP accurately attends to disease-infected regions, confirming that the model learns semantically meaningful and visually grounded representations. These insights highlight that progressive adaptation not only improves performance, but also improves transparency and reliability, which are crucial factors for agricultural applications. From a practical standpoint, PLA-CLIP achieves a strong balance between accuracy and computational efficiency, reaching an inference time of 2.65 ms while maintaining a manageable model size. This efficiency, combined with the interpretability of the model, establishes PLA-CLIP as a strong foundation to advance intelligent automation and decision-support systems in agricultural engineering. Future research will focus on developing adaptive layer-unfreezing strategies, extending the framework to various crop types, and integrating temporal and environmental data for predictive analysis and large-scale monitoring within automated agricultural systems.

ACKNOWLEDGEMENTS

This work was supported in part by the Khalifa University Center for Autonomous and Robotic Systems (KUCARS) under Award RC1-2018-KUCARS and in part by Silal IO, UAE Leading Food and Technology Company, which is jointly funded by Silal Innovation, part of Silal and a State-of-the-Art R&D Center in Al-Ain.

REFERENCES

[1] A. Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, and D. P. Hughes, "Deep learning for image-based cassava disease detection," *Frontiers in Plant Science*, vol. 8, p. 1852, 2017.

[2] J. P. Legg, S. C. Jeremiah, H. M. Obiero, M. N. Maruthi, I. Ndyetabula, G. Okao-Okuja, H. Bouwmeester, S. Bigirimana, W. Tata-Hangy, G. Gashaka, *et al.*, "Cassava mosaic geminiviruses in africa," *Plant Disease*, vol. 90, no. 12, pp. 1486–1498, 2006.

[3] D. Velayudhan, M. Shafay, S. G. Khawaja, N. Werghi, and T. Hassan, "Semi-supervised segmentation-driven classification pipeline for grading cassava leaf diseases," in *2024 International Conference on Engineering and Emerging Technologies (ICEET)*, pp. 1–6, 2024.

[4] J. Liu, F. Lv, and B. Liu, "Identification method of sunflower leaf disease based on sift point," *Journal of Image and Graphics*, vol. 7, no. 2, pp. 64–67, 2019.

[5] K. Saminathan, B. Sowmiya, and M. C. Devi, "Multiclass classification of paddy leaf diseases using random forest classifier," *Journal of Image and Graphics*, vol. 11, no. 2, pp. 195–203, 2023.

[6] M. Shafay, T. Hassan, A. Ahmed, D. Velayudhan, J. Dias, and N. Werghi, "Programmable broad learning system to detect concealed and imbalanced baggage threats," in *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, pp. 1–6, IEEE, 2022.

[7] M. Alansari, A. Ahmed, K. Alnuaimi, D. Velayudhan, T. Hassan, S. Javed, M. Bennamoun, and N. Werghi, "Multi-scale hierarchical contour framework for detecting cluttered threats in baggage security," *IEEE Access*, vol. 12, pp. 77454–77467, 2024.

[8] A. Ahmed, M. Alansari, K. Alnuaimi, D. Velayudhan, T. Hassan, and N. Werghi, "Detection transformer framework for recognition of heavily occluded suspicious objects," in *2023 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, pp. 1–6, IEEE, 2023.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, pp. 8748–8763, PMLR, 2021.

[10] M. F. Nasir, M. U. Rehman, and I. Hussain, "Vision-language models for zero-shot weed detection and visual reasoning in uav-based precision agriculture," *Frontiers in Plant Science*, vol. 16, p. 1735096, 2026.

[11] M. U. Din, W. Akram, A. B. Bakht, Y. Dong, and I. Hussain, "Maritime mission planning for unmanned surface vessel using large language model," in *2025 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*, pp. 1–6, April 2025.

[12] J. Zhang *et al.*, "Maianet: signal modulation in cassava leaf disease classification," *Computers and Electronics in Agriculture*, vol. 225, p. 109351, 2024.

[13] A. Tewari and S. Kumari, "A lightweight model incorporating channel attention mechanisms for cassava disease detection," *Expert Systems with Applications*, vol. 215, p. 119332, 2024.

[14] J. Liu *et al.*, "Research on cassava disease classification using the multi-scale fusion model based on efficientnet and attention mechanism," *PLoS One*, vol. 18, no. 1, p. e0278734, 2023.

[15] M. Owais, M. Shafay, M. Zubair, S. Mohammed, L. Seneviratne, and I. Hussain, "Agrivision: A benchmark dataset for advancing real-world robotic vision in densely fruited blueberry crop," *Scientific Data*, vol. 12, p. 2014, Dec 2025.

[16] M. Shafay, T. Hassan, M. Owais, I. Hussain, S. G. Khawaja, L. Seneviratne, and N. Werghi, "Recent advances in plant disease detection: challenges and opportunities," *Plant Methods*, vol. 21, p. 140, Oct 2025.

[17] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[18] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," in *International Journal of Computer Vision*, 2021.

[19] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *European conference on computer vision*, pp. 493–510, Springer, 2022.

[20] M. W. Gondal, J. Gast, I. A. Ruiz, R. Droste, T. Macri, S. Kumar, and L. Staudigl, "Domain aligned clip for few-shot classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5721–5730, 2024.

[21] J. Silva-Rodriguez, J. Dolz, and I. Ben Ayed, "A closer look at the few-shot adaptation of large vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18647–18656, 2024.

[22] ErnestMwebaze, J. Mostipak, Joyce, J. Elliott, and S. Dane, "Cassava leaf disease classification." https://kaggle.com/competitions/cassava-leaf-disease-classification, 2020. Kaggle.

[23] F. J. Tusubira, J. Nakatumba-Nabende, C. Babirye, G. Okao-Okujja, C. Mutebi, B. W. Mugalu, D. Nabagereka, and G. Namanya, "Makerere University Cassava Image Dataset," 2022.

[24] L. G. Divyanth, P. Soni, and R. Machavaram, "Cassava leaf disease dataset," 2021.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[28] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[29] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 2FinalPresentationPaperSE103ICCAE2026CameraReadyVersion.pdf