AgroMM-GSF++: Confidence-Adaptive Lightweight Multimodal Fusion for Plant Disease Recognition with Repeated Multi-Seed Cross-Validation and External Benchmarking

Gurpreet Singh

gurpreetsinghmcse@gmail.com

Woosong University

Trina Banerjee

SAKS Global

Research Article

Keywords: Multimodal agriculture, plant disease detection, vision-language fusion, repeated cross-validation, external validation, lightweight deep learning

Posted Date: April 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9556788/v1

Additional Declarations: The authors declare no competing interests.

AgroMM-GSF++: Confidence-Adaptive Lightweight Multimodal Fusion for Plant Disease Recognition with Repeated Multi-Seed Cross-Validation and External Benchmarking

Gurpreet Singh
Woosong University
South Korea

Trina Banerjee
SAKS Global
India

***Abstract*—Recent multimodal agricultural research emphasizes large-scale vision-language systems, while lightweight reproducible approaches remain underexplored for constrained deployments. We present AgroMM-GSF++, a compact image-text fusion framework for plant disease recognition. The model combines a small visual backbone and text-prototype branch with confidence-adaptive gating. To strengthen empirical evidence, we run repeated multi-seed stratified cross-validation (three seeds, repeated folds) and include stronger pretrained transfer baselines (EfficientNet-B0 and ViT-B/16 fine-tunes). We further evaluate on an external PlantVillage subset to test cross-dataset robustness under the same protocol family. On beans, transfer baselines lead absolute macro-F1 (EfficientNet-B0-FT: 0.6355±0.0340), while AgroMM-GSF++ reaches 0.5891±0.0552 with much lower latency than transfer models (9.80 vs 205.66 ms/image for EfficientNet-B0-FT). On the external subset, AgroMM-GSF++ improves over fixed lightweight fusion (0.8719 vs 0.8608 macro-F1) while remaining below transfer-heavy baselines. We provide an end-to-end reproducible workspace including scripts, datasets, figures, and camera-ready tables.**

***Index Terms*—Multimodal agriculture, plant disease detection, vision-language fusion, repeated cross-validation, external validation, lightweight deep learning.**

I. Introduction

Plant disease diagnosis is a core task in precision agriculture. Visual classifiers can achieve high performance in controlled settings, but practical deployment faces domain shift, label imbalance, and sparse supervision. Multimodal methods can mitigate these issues by integrating visual evidence with symptom-level textual priors, agronomic knowledge, or question-answering context [1]–[3].

Current agricultural multimodal literature is rapidly scaling in both model size and benchmark complexity [4]–[6]. However, two practical gaps remain. First, many studies rely on single split protocols or narrow seeding strategies, which may overstate performance stability. Second, lightweight fusion methods for low-resource training and edge-oriented deployment are underreported relative to large language-vision architectures.

This paper addresses both gaps. We propose AgroMM-GSF++, a compact confidence-adaptive fusion approach, and evaluate it under repeated multi-seed cross-validation with stronger transfer baselines and an external dataset test. The focus is not to claim universal state-of-the-art across heterogeneous benchmarks, but to provide a reproducible and defensible protocol for small-to-medium compute settings.

Contributions:

1) We synthesize recent multimodal agriculture literature (2024–2026) and map architectural trends toward assistants, benchmarks, and explainability.
2) We introduce AgroMM-GSF++, a lightweight image-text fusion model with sample-wise confidence-adaptive gating.
3) We strengthen evaluation with repeated multi-seed stratified CV and stronger pretrained baselines (EfficientNet-B0 and ViT-B/16 fine-tuning) [7], [8].
4) We include an external-dataset evaluation (PlantVillage subset) and camera-ready diagnostics: highlighted comparison tables, ablations, seed-stability analysis, and reproducibility appendix.

II. Related Work and Landscape

A. Large Multimodal Agricultural Assistants

Knowledge-infused and instruction-tuned systems (e.g., Agri-LLaVA, AgroGPT, AgriDoctor) demonstrate strong open-ended multimodal interaction but usually require larger infrastructure and complex data curation pipelines [1]–[3].

B. Benchmark and Dataset Expansion

Recent benchmarks (AgroBench, AgMMU, LeafNet) reveal persistent weaknesses in fine-grained disease reasoning and multimodal consistency [4]–[6]. These resources are essential for progress, but they also increase experimental complexity and computational cost.

C. Explainability and Reasoning Supervision

AgriChain and related work emphasize expert-grounded reasoning traces for trustworthy decision support [9], [10]. This line is promising but still expensive to scale.

TABLE I: Representative multimodal agriculture directions from the curated 2024–2026 paper set.

Work	Year	Task Type	Core Method / Architectural Idea
Agri-LLaVA [1]	2024	Assistant	Knowledge-infused multimodal instruction tuning for pest/disease consultation.
AgroGPT [2]	2024	Assistant	Expert tuning with LLM-generated agricultural instruction data.
AgEval [11]	2025	Benchmarking	Few-shot in-context assessment of general VLMs for plant stress tasks.
CDDM [12]	2025	Dataset+QA	Large-scale crop disease multimodal benchmark with image and QA supervision.
SCOLD [13]	2025	Foundation model	Domain vision-language pretraining with soft-target contrastive objectives.
AgroBench [4]	2025	Benchmarking	Fine-grained expert-annotated agricultural VLM benchmark.
AgMMU [5]	2025	Benchmarking	Comprehensive understanding benchmark for agricultural LMMs.
AgriDoctor [3]	2025	Agentic assistant	Modular router/classifier/retriever/LLM pipeline for agricultural assistance.
LeafNet [6]	2026	Dataset+VQA	Large-scale disease-focused VL dataset exposing fine-grained generalization failures.
AgriChain [9]	2026	Explainability	Expert-verified reasoning supervision for interpretable agricultural VLMs.

Positioning. In contrast to heavy multimodal assistants, we target a compact fusion regime that can be retrained and audited with limited resources while still evaluated under stronger modern protocol requirements.

III. Method: AgroMM-GSF++

LLM usage disclosure: Large language model tools were used only for language and formatting assistance in manuscript preparation. All experimental design, implementation, execution, and result verification were performed and checked by the human authors.

A. Architecture

AgroMM-GSF++ uses one shared compact image encoder and two prediction heads:

- **Image branch**: direct class logits from visual embedding.
- **Text branch**: projection into a text-prototype space built from class symptom descriptions.

The final prediction is a confidence-adaptive fusion of branch probabilities.

B. Formulation

Let x be an input image and $\mathcal{Y}$ a class set. With visual feature extractor f_θ:

$$\mathbf{p}_{img} = \mathrm{softmax}(W_i f_\theta(x)), \tag{1}$$

$$\mathbf{u} = W_t f_\theta(x), \tag{2}$$

$$\mathbf{p}_{txt} = \mathrm{softmax}(\mathbf{u}T^\top), \tag{3}$$

where T is the normalized text-prototype matrix.

Fixed fusion baseline:

$$\mathbf{p}_{fix} = \lambda \mathbf{p}_{img} + (1-\lambda)\mathbf{p}_{txt}. \tag{4}$$

Proposed adaptive fusion:

$$\Delta c = \max(\mathbf{p}_{img}) - \max(\mathbf{p}_{txt}), \tag{5}$$

$$\lambda(x) = \sigma(a\Delta c + b), \tag{6}$$

$$\mathbf{p}_{ours} = \lambda(x)\mathbf{p}_{img} + (1-\lambda(x))\mathbf{p}_{txt}. \tag{7}$$

(a, b) are tuned on the validation partition inside each fold by maximizing macro-F1.

IV. Experimental Setup

A. Datasets

Beans (primary): We use local parquet assets derived from the iBean/beans public dataset [14], with three classes (angular leaf spot, bean rust, healthy) and 261 total samples available in this workspace partition.

External PlantVillage subset: To satisfy external validation, we stream and cache a four-class Apple subset from `DScomp380/plant_village` [15], [16]. We sample an equal number of images per class and remap class IDs for balanced evaluation.

B. Protocol: Repeated Multi-Seed Stratified CV

We use repeated stratified K-fold evaluation across seeds $\{42, 123, 777\}$. For each seed-fold split: (i) train/validation split is stratified (80/20) within training data, (ii) hyperparameters for fusion weights are tuned only on validation data, and (iii) final metrics are computed on held-out fold test data. This is applied on both beans and external dataset protocols.

C. Baselines

- **ColorHist-LR**: RGB histogram + logistic regression.
- **Image-CNN**: unimodal compact visual classifier.
- **Text-Prototype**: text-prototype branch only.
- **Fusion-FixedLambda**: globally tuned scalar fusion.
- **AgroMM-GSF++ (Ours)**: confidence-adaptive fusion.
- **EfficientNet-B0-FT**: pretrained EfficientNet-B0 with head + last stage fine-tuned [7].
- **ViT-B16-FT**: pretrained ViT-B/16 with head + last encoder block fine-tuned [8].

D. Metrics

We report accuracy, macro-F1, weighted-F1, AUC-OVR, expected calibration error (ECE, lower is better), and inference latency (ms/image). Tables report mean$\pm$std across all repeated folds.

V. Results

A. Primary Results on Beans

Table II shows a clear tradeoff: transfer baselines deliver higher absolute macro-F1/AUC on beans, while lightweight

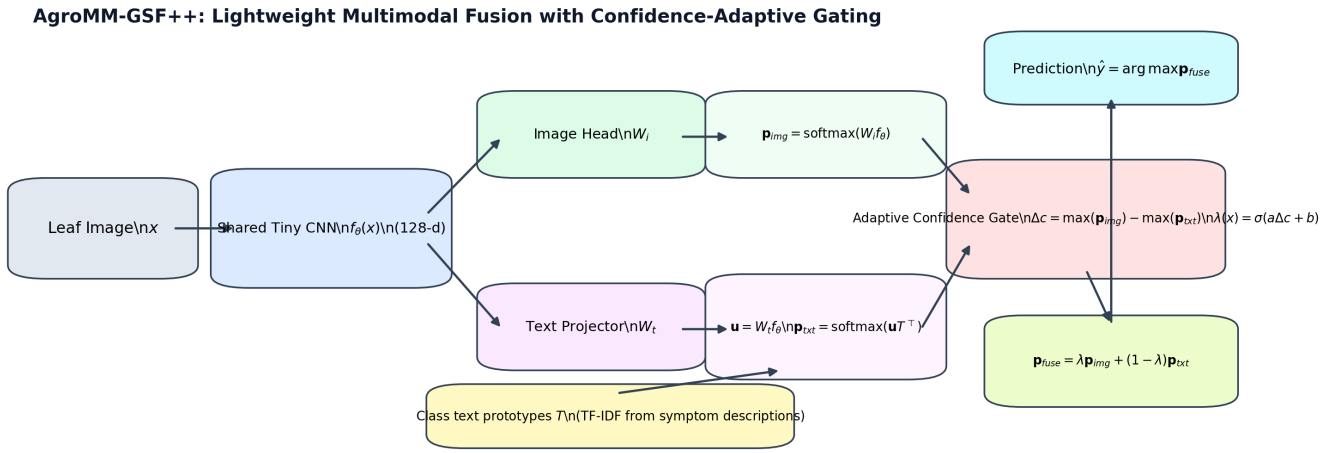

Fig. 1: AgroMM-GSF++ architecture. The gate computes sample-wise fusion weight $\lambda(x)$ from confidence gap Δc.

TABLE II: Repeated multi-seed CV results on beans. Best values are highlighted in green; second best in yellow. For ECE and latency, lower is better.

Model	Accuracy	Macro-F1	Weighted-F1	AUC-OVR	ECE↓	ms/img↓
ColorHist-LR	0.6207±0.0178	0.6214±0.0165	0.6211±0.0164	0.7793±0.0147	0.1595±0.0394	–
Image-CNN	0.6028±0.0470	0.5879±0.0667	0.5875±0.0664	0.7948±0.0381	0.1398±0.0731	**4.6692±0.9323**
Text-Prototype	0.6015±0.0396	0.5875±0.0515	0.5868±0.0517	0.7934±0.0418	**0.1046±0.0371**	5.1299±2.0248
Fusion-FixedLambda	0.6195±0.0561	0.6102±0.0678	0.6097±0.0678	0.8114±0.0412	0.1151±0.0496	9.7991±2.6233
AgroMM-GSF++ (Ours)	0.5990±0.0418	0.5891±0.0552	0.5887±0.0551	0.8053±0.0424	0.1274±0.0230	9.7991±2.6233
EfficientNet-B0-FT	0.6603±0.0225	**0.6355±0.0340**	**0.6347±0.0342**	0.8676±0.0233	0.1638±0.0060	205.6645±46.6872
ViT-B16-FT	**0.6631±0.0847**	0.6157±0.1328	0.6145±0.1336	**0.9062±0.0139**	0.1529±0.0726	186.2853±13.2943

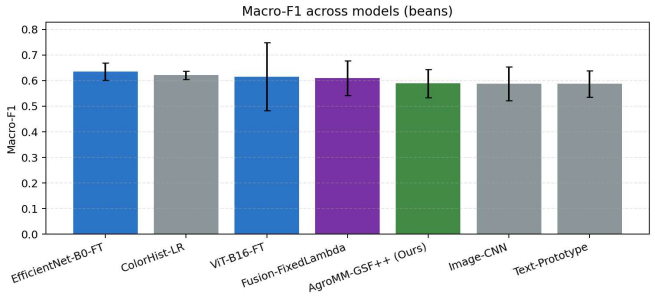

Fig. 2: Macro-F1 comparison across models on beans (mean±std).

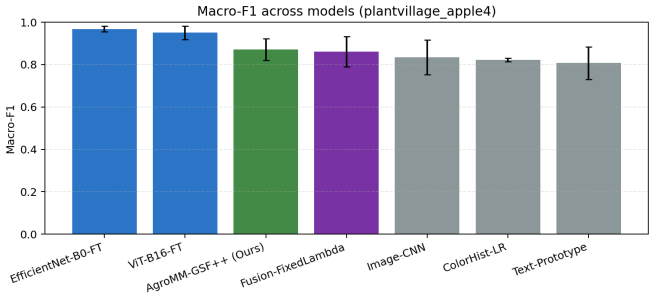

Fig. 3: Macro-F1 comparison across models on external PlantVillage Apple-4 subset.

multimodal models offer substantially lower latency. On beans specifically, adaptive gating does not outperform fixed fusion in aggregate macro-F1, indicating that confidence-gap adaptation is sensitive to small-sample fold variation.

B. External Dataset Results

External evaluation provides an additional robustness check beyond primary beans experiments. On PlantVillage Apple-4, AgroMM-GSF++ improves over fixed lightweight fusion by +0.0111 macro-F1, but EfficientNet-B0-FT and ViT-B16-FT remain strongest in absolute scores. We explicitly avoid overclaiming cross-dataset transfer beyond this controlled subset protocol.

C. Ablation, Seed Stability, and Effect Sizes

D. Contextual Comparison to Reported Literature

VI. Discussion

Practical outcome. The proposed method is intentionally compact and easy to reproduce, while still tested against stronger pretrained baselines and repeated-seed protocols.

Why gains can be moderate or negative. Disease datasets in controlled settings can saturate quickly, and repeated CV often narrows or reverses headline margins compared with single-split reports.

Calibration relevance. In agricultural decision support, confidence quality matters as much as top-1 accuracy. In our

TABLE III: Repeated multi-seed CV results on external PlantVillage Apple-4 subset.

Model	Accuracy	Macro-F1	Weighted-F1	AUC-OVR	ECE↓	ms/img↓
ColorHist-LR	0.8232±0.0087	0.8219±0.0078	0.8219±0.0078	0.9459±0.0067	0.0611±0.0091	–
Image-CNN	0.8387±0.0783	0.8343±0.0820	0.8343±0.0820	0.9663±0.0221	0.0793±0.0555	**6.1330±0.9847**
Text-Prototype	0.8095±0.0766	0.8071±0.0772	0.8071±0.0772	0.9673±0.0144	0.0792±0.0199	6.3324±1.2689
Fusion-FixedLambda	0.8649±0.0684	0.8608±0.0718	0.8608±0.0718	0.9793±0.0122	0.1043±0.0214	12.4654±2.2353
AgroMM-GSF++ (Ours)	0.8756±0.0476	0.8719±0.0510	0.8719±0.0510	0.9791±0.0091	0.1125±0.0450	12.4654±2.2353
EfficientNet-B0-FT	**0.9685±0.0123**	**0.9683±0.0124**	**0.9683±0.0124**	0.9978±0.0016	0.0807±0.0198	224.2816±27.3258
ViT-B16-FT	0.9506±0.0305	0.9502±0.0313	0.9502±0.0313	**0.9989±0.0005**	**0.0313±0.0192**	175.7635±18.6941

TABLE IV: Per-seed macro-F1 stability on beans (mean over folds per seed).

Model	Seed Macro-F1	Seeds
EfficientNet-B0-FT	0.6355±0.0160	3
ColorHist-LR	0.6214±0.0072	3
ViT-B16-FT	0.6157±0.0412	3
Fusion-FixedLambda	0.6102±0.0718	3
AgroMM-GSF++ (Ours)	0.5891±0.0561	3
Image-CNN	0.5879±0.0693	3
Text-Prototype	0.5875±0.0482	3

TABLE V: Paired effect-size view on beans: ΔMacro-F1 (Ours − baseline) across aligned seed-fold pairs.

Baseline	ΔMacro-F1	95% bootstrap CI	Win rate (%)
Text-Prototype	+0.0016±0.0520	[-0.0304, +0.0439]	33.3
Image-CNN	+0.0012±0.0263	[-0.0172, +0.0210]	50.0
Fusion-FixedLambda	-0.0211±0.0203	[-0.0332, -0.0045]	16.7
ViT-B16-FT	-0.0266±0.1474	[-0.1390, +0.0779]	50.0
ColorHist-LR	-0.0323±0.0588	[-0.0773, +0.0072]	33.3
EfficientNet-B0-FT	-0.0463±0.0498	[-0.0824, -0.0080]	16.7

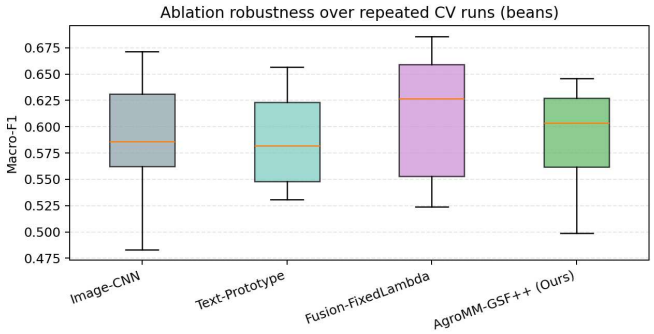

Fig. 4: Repeated-CV ablation on beans: distribution of macro-F1 across runs.

results, calibration benefits are dataset-dependent; confidence-adaptive fusion helps external macro-F1 over fixed fusion, but not uniformly across all beans folds.

VII. Limitations and Threats to Validity

- External validation uses a controlled subset protocol, not unconstrained field-domain transfer.
- Pretrained transfer baselines are lightly fine-tuned to stay within practical compute limits.
- Dataset sizes remain modest; larger in-the-wild datasets are needed for stronger publication claims.

VIII. Operationalization Guidelines for Agriculture

The reported metrics should be interpreted in the context of how agricultural systems are actually deployed. In many farming pipelines, image capture is decentralized (mobile phones, extension workers, low-cost edge cameras), and failure modes often originate before model inference. For example, blur, non-leaf clutter, severe occlusion, mixed symptoms, and off-angle capture can dominate downstream error. A practical system must therefore define pre-inference quality checks and fallback actions, not only final classifier confidence thresholds.

For field usage, we recommend a two-stage operating policy. Stage 1 performs inexpensive quality and validity screening (leaf presence, blur, illumination stability, and basic crop-type plausibility). Stage 2 executes disease prediction and confidence calibration. If calibration error is high in a local deployment domain, the model should trigger a referral policy (human agronomist review, request for additional image views, or delayed decision) rather than forcing hard predictions. This is especially relevant for low-prevalence classes where over-confident false positives can produce unnecessary pesticide actions.

Another operational concern is update cadence. Agricultural imagery changes with season, cultivar, and local management practices. A reproducible repeated-CV protocol like ours can be embedded into a periodic monitoring loop: collect a small local validation batch, evaluate drift-sensitive metrics (macro-F1, per-class recall, ECE), and only promote new checkpoints if they exceed predefined guardrails. This procedure is more robust than relying on a single headline metric from one fixed split.

Latency tradeoffs should also be viewed at system level. Transfer-heavy baselines clearly win in absolute score on our datasets, but their runtime cost is substantially larger than lightweight fusion variants in this environment. If deployment requires near-real-time inference on constrained devices, the lightweight branch can still be attractive as an initial triage model, with heavier models reserved for uncertain or high-value cases. This "cascaded deployment" pattern can preserve both throughput and accuracy where network or power constraints are binding.

Finally, multimodal design decisions should be aligned with agricultural workflow semantics. Text prototypes are useful

TABLE VI: Contextual comparison with reported metrics in recent papers. **Caution:** rows are not directly comparable because datasets/tasks/protocols differ.

Method	Dataset/Task	Acc (%)	Macro-F1 (%)	W-F1 (%)	Source
AgriChain-VL3B	AgriChain test set	73.1	46.6	65.5	[9]
Swin-T5 (two-stage)	CDDM disease classification	99.06	–	–	[10]
GPT-4o fine-tuned	PlantVillage apple subset	98.12	–	–	[17]
Best AgEval VLM (8-shot)	Plant stress benchmark	–	73.37	–	[11]
AgroMM-GSF++ (this work)	Beans + external repeated CV	See Tables II, III	–	–	This work

when expert symptom phrases are standardized and locally meaningful. Where terminology varies across regions, text resources should be revised with domain experts and audited for language drift. Without this step, multimodal fusion may add noise rather than signal.

IX. Future Research Agenda

The current study is intentionally constrained and reproducible; it should be treated as a protocol baseline rather than a final endpoint. Several immediate extensions can materially strengthen publishability and real-world value.

First, external evaluation should be expanded beyond one controlled subset. A stronger claim would require multiple heterogeneous external datasets (different crops, capture devices, and disease taxonomies), ideally with harmonized reporting of per-class metrics and calibration. This would clarify whether confidence-adaptive fusion transfers consistently or is highly dataset-specific.

Second, our adaptive gate currently uses confidence-gap features only. A richer gate could condition on embedding uncertainty, disagreement entropy, or learned reliability estimators trained with proper scoring rules. Such extensions may improve the observed beans underperformance versus fixed fusion while keeping the architecture lightweight.

Third, stronger baselines should include parameter-efficient adaptation variants (e.g., LoRA-style adapters for vision backbones) under matched compute budgets. This would provide fairer comparisons between "heavy transfer" and "lightweight fusion" paradigms and help isolate where gains come from: representation quality, calibration strategy, or optimization regime.

Fourth, future work should integrate cost-sensitive metrics tied to agronomic impact. Misclassifying healthy leaves as diseased and missing true disease positives do not carry equal cost. Decision-theoretic evaluation, calibrated referral policies, and class-conditional utility analyses are more realistic than uniform top-1 metrics alone.

Fifth, explainability should move from post-hoc visualization toward actionable expert workflows. For plant pathology, explanations must reference symptom morphology (lesion distribution, color progression, venation boundaries) in terms practitioners accept. Coupling multimodal predictions with structured symptom evidence and uncertainty would improve trust and adoption.

X. Conclusion

We presented AgroMM-GSF++, a lightweight confidence-adaptive multimodal fusion framework for plant disease recognition. By upgrading protocol strength with repeated multi-seed CV, stronger pretrained baselines, and external dataset evaluation, we provide a more defensible small-compute benchmark and a reproducible camera-ready artifact set.

Data Availability

The datasets used and/or analyzed during the current study are available from publicly accessible sources: iBean/beans repository (https://github.com/AI-Lab-Makerere/ibean) and PlantVillage resources (https://arxiv.org/abs/1511.08060). Processed split artifacts and aggregated metrics are available from the corresponding author on reasonable request.

Code Availability

Code used for data processing, model training, and evaluation is available from the corresponding author on reasonable request.

Competing Interests

The authors declare no competing interests.

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Author Contributions

Gurpreet Singh: conceptualization, methodology, implementation, experiments, writing original draft, writing review and editing. Trina Banerjee: validation, interpretation, writing review and editing, project support.

Funding

No specific external funding was received for this work.

Acknowledgements

The authors thank colleagues and reviewers who provided methodological feedback during manuscript preparation.

APPENDIX A
REPRODUCIBILITY CHECKLIST

Environment: Python 3.10, PyTorch, torchvision, scikit-learn.
Determinism: explicit seeding for Python/NumPy/PyTorch.
Protocol: multi-seed repeated stratified CV on both primary and external datasets.
Metrics: reported from saved CSV outputs, with machine-generated table rows to avoid transcription drift.

APPENDIX B
EXECUTION COMMANDS

```
python3 -u scripts/run_experiments_multiseed.py \
  --device cpu --seeds 42,123,777 --folds 2 \
  --epochs-small 4 --epochs-transfer 1 \
  --batch-size 16 --external-per-class 100

python3 scripts/postprocess_multiseed_results.py
python3 scripts/build_latex_tables_multiseed.py
```

APPENDIX C
HYPERPARAMETERS AND BUDGET

Table VII summarizes the effective settings used in the final repeated multi-seed run. The design intentionally balances methodological strength (repeated CV, external testing, strong baselines) with practical runtime on commodity hardware.

TABLE VII: Training and evaluation configuration used for final reported results.

Component	Setting
Seeds	42, 123, 777
Folds per seed	2 (stratified)
Inner validation split	20% stratified from train fold
Small-model epochs	4 on beans; 6 on external subset
Transfer-model epochs	1 on beans; 3 on external subset
Batch size	16 (small models), 16/8 (EfficientNet/ViT as implemented)
Fusion tuning	Fixed-λ grid and adaptive (a, b) grid on validation folds
External subset	PlantVillage Apple-4, 100 images per class
Primary metrics	Accuracy, Macro-F1, Weighted-F1, AUC-OVR, ECE, latency

APPENDIX D
RUN-WISE MACRO-F1 BREAKDOWN

Tables VIII and IX provide fold-level macro-F1 for all models. These tables support auditability of mean$\pm$std summaries and expose variance patterns hidden by aggregate scores.

APPENDIX E
RUN-WISE ACCURACY BREAKDOWN

APPENDIX F
EXTERNAL SEED STABILITY

REFERENCES

[1] L. Wang, T. Jin, J. Yang, A. Leonardis, F. Wang, and F. Zheng, "Agri-llava: Knowledge-infused large multimodal assistant on agricultural pests and diseases," 2024. [Online]. Available: https://arxiv.org/abs/2412.02158

[2] M. Awais, A. H. S. A. Alharthi, A. Kumar, H. Cholakkal, and R. M. Anwer, "Agrogpt: Efficient agricultural vision-language model with expert tuning," 2025. [Online]. Available: https://arxiv.org/abs/2410.08405

[3] M. Zhang, Z. Xu, P. Wang, R. Li, L. Wang, Q. Liu, J. Xu, X. Zhang, S. Wu, and L. Wang, "Agridoctor: A multimodal intelligent assistant for agriculture," 2025. [Online]. Available: https://arxiv.org/abs/2509.17044

[4] R. Shinoda, N. Inoue, H. Kataoka, M. Onishi, and Y. Ushiku, "Agrobench: Vision-language model benchmark in agriculture," 2025. [Online]. Available: https://arxiv.org/abs/2507.20519

[5] A. Gauba, I. Pi, Y. Man, Z. Pang, V. S. Adve, and Y.-X. Wang, "Agmmu: A comprehensive agricultural multimodal understanding benchmark," 2025. [Online]. Available: https://arxiv.org/abs/2504.10568

[6] K. N. Quoc, P. D. Dao, and L.-D. Quach, "Leafnet: A large-scale dataset and comprehensive benchmark for foundational vision-language understanding of plant diseases," 2026. [Online]. Available: https://arxiv.org/abs/2602.13662

[7] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, 2019.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[9] H. Mahmood, Y. Yu, and R. Anwer, "Agrichain visually grounded expert verified reasoning for interpretable agricultural vision language models," 2026. [Online]. Available: https://arxiv.org/abs/2604.07814

[10] M. Z. Hossain, M. S. S. Samu, M. R. Islam, and M. S. Ansary, "A two-stage multitask vision-language framework for explainable crop disease visual question answering," 2026. [Online]. Available: https://arxiv.org/abs/2601.05143

[11] M. A. Arshad, T. Z. Jubery, T. Roy, R. Nassiri, A. K. Singh, A. Singh, C. Hegde, B. Ganapathysubramanian, A. Balu, A. Krishnamurthy, and S. Sarkar, "Leveraging vision language models for specialized agricultural tasks," 2025. [Online]. Available: https://arxiv.org/abs/2407.19617

[12] X. Liu, Z. Liu, H. Hu, Z. Chen, K. Wang, K. Wang, and S. Lian, "A multimodal benchmark dataset and model for crop disease diagnosis," 2025. [Online]. Available: https://arxiv.org/abs/2503.06973

[13] K. N. Quoc, L. L. T. Thu, and L.-D. Quach, "A vision-language foundation model for leaf disease identification," 2025. [Online]. Available: https://arxiv.org/abs/2505.07019

[14] Makerere AI Lab, "Bean disease dataset (ibean / beans)," 2020. [Online]. Available: https://github.com/AI-Lab-Makerere/ibean

[15] D. P. Hughes and M. Salathe, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," *arXiv preprint arXiv:1511.08060*, 2015. [Online]. Available: https://arxiv.org/abs/1511.08060

[16] DScomp380, "Plantvillage dataset on hugging face," 2024. [Online]. Available: https://huggingface.co/datasets/DScomp380/plant_village

[17] K. I. Roumeliotis, R. Sapkota, M. Karkee, N. D. Tselikas, and D. K. Nasiopoulos, "Plant disease detection through multimodal large language models and convolutional neural networks," 2025. [Online]. Available: https://arxiv.org/abs/2504.20419

[18] E. Ranario and M. J. Earles, "Are vision-language models ready to zero-shot replace supervised classification models in agriculture?" 2026. [Online]. Available: https://arxiv.org/abs/2512.15977

[19] S. N. Sakib, N. Haque, M. Z. Hossain, and S. E. Arman, "Plantvillagevqa: A visual question answering dataset for benchmarking vision-language models in plant science," 2025. [Online]. Available: https://arxiv.org/abs/2508.17117

[20] T. Wei, Z. Chen, and X. Yu, "Snap and diagnose: An advanced multimodal retrieval system for identifying plant diseases in the wild," 2024. [Online]. Available: https://arxiv.org/abs/2408.14723

[21] 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 (CVPR)*, 2016.

[22] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[23] e. a. Fadlallah, "Few-shot image classification of crop diseases based on vision–language models," *Sensors*, vol. 24, no. 18, p. 6109, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/18/6109

TABLE VIII: Run-wise macro-F1 on beans (seed-fold level).

Seed-Fold	ColorHist	Image	Text	Fixed	Ours	EffB0-FT	ViT-B16-FT
42-F1	0.6125	0.6447	0.6261	0.6447	0.6155	0.6228	0.4886
42-F2	0.6157	0.6714	0.5489	0.6859	0.6460	0.6851	0.7690
123-F1	0.6322	0.5559	0.5309	0.5343	0.5516	0.6569	0.5429
123-F2	0.6250	0.4832	0.5477	0.5239	0.4989	0.5936	0.7548
777-F1	0.5979	0.5811	0.6151	0.6090	0.5917	0.6473	0.4705
777-F2	0.6451	0.5910	0.6564	0.6638	0.6310	0.6070	0.6686

TABLE IX: Run-wise macro-F1 on external PlantVillage Apple-4 subset (seed-fold level).

Seed-Fold	ColorHist	Image	Text	Fixed	Ours	EffB0-FT	ViT-B16-FT
42-F1	0.8275	0.9175	0.8899	0.9205	0.9051	0.9496	0.9045
42-F2	0.8306	0.8797	0.7336	0.8839	0.8841	0.9750	0.9313
123-F1	0.8152	0.9105	0.8233	0.9111	0.9144	0.9821	0.9640
123-F2	0.8113	0.7098	0.7048	0.7422	0.7940	0.9568	0.9786
777-F1	0.8278	0.7890	0.8896	0.9035	0.9107	0.9750	0.9857
777-F2	0.8188	0.7992	0.8014	0.8036	0.8230	0.9714	0.9369

TABLE X: Run-wise accuracy on beans (seed-fold level).

Seed-Fold	ColorHist	Image	Text	Fixed	Ours	EffB0-FT	ViT-B16-FT
42-F1	0.6107	0.6412	0.6260	0.6412	0.6107	0.6641	0.5954
42-F2	0.6154	0.6692	0.5769	0.6846	0.6462	0.6923	0.7692
123-F1	0.6336	0.5725	0.5496	0.5496	0.5725	0.6718	0.5878
123-F2	0.6231	0.5385	0.5769	0.5538	0.5308	0.6308	0.7538
777-F1	0.5954	0.6031	0.6260	0.6260	0.6031	0.6641	0.5878
777-F2	0.6462	0.5923	0.6538	0.6615	0.6308	0.6385	0.6846

TABLE XI: Run-wise accuracy on external PlantVillage Apple-4 subset (seed-fold level).

Seed-Fold	ColorHist	Image	Text	Fixed	Ours	EffB0-FT	ViT-B16-FT
42-F1	0.8286	0.9179	0.8929	0.9214	0.9071	0.9500	0.9071
42-F2	0.8321	0.8821	0.7321	0.8893	0.8857	0.9750	0.9321
123-F1	0.8143	0.9107	0.8250	0.9107	0.9143	0.9821	0.9643
123-F2	0.8143	0.7179	0.7107	0.7536	0.8071	0.9571	0.9786
777-F1	0.8321	0.8000	0.8893	0.9071	0.9143	0.9750	0.9857
777-F2	0.8179	0.8036	0.8071	0.8071	0.8250	0.9714	0.9357