AgriPerceiver: A Parameter-Efficient Vision-Language Model for Structured Macroscopic Crop Phenotyping

Vatsal Khanna [1] **Davinder Singh** [1]

Abstract

Modern agriculture suffers from systemic crop yield instability, with plant pathogens contributing to 20-40% of global yield losses. Adapting generalist Vision-Language Models (VLMs) to produce structured, actionable phenotyping from crop images, creates a scalability crisis at the intersection of computer vision and agricultural life sciences. We present **AgriPerceiver**, a lightweight VLM that frames this challenge as *structured report generation*: given a single leaf photograph, the model produces a schema-compliant JSON diagnostic report detailing disease identity, pathology type, severity score, symptom characterisation, and actionable steps. To process high-resolution visual imagery, the input is spatially decomposed to preserve fine-grained pathological features that are typically lost during standard resizing. Our central contribution is a *perception bridge* that mitigates visual token explosion by compressing visual tokens into just 128 learned latents ($28.5\times$ reduction) while critically preserving learned tile-position embeddings for spatial grounding. We employ a two-stage training curriculum: the bridge ($\sim$391M parameters) is first aligned, followed by LoRA specialization on Gemma-3-labelled structured annotations. Both the vision and language backbones remain strictly frozen throughout training to maintain parameter efficiency. Evaluated across nine metrics, AgriPerceiver achieves a composite score of **0.810** and 99.7% schema compliance on a held-out test set, demonstrating the viability of parameter-efficient domain specialization in life sciences AI for structured knowledge extraction.

[*]Equal contribution [1]Department of Electronics and Communication Engineering, Dr. B R Ambedkar National Institute of Technology Jalandhar, Punjab, India. Correspondence to: Vatsal Khanna <vatsalk.ec.23@nitj.ac.in>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Pathogen-induced crop losses account for 20–40% of global agricultural output annually (Savary et al., 2019), making plant disease diagnosis one of the highest-impact problems in life sciences and food security. Conventional diagnostic workflows require trained agronomists to inspect field samples and produce structured reports that cover disease type, severity, observed symptoms, and treatment recommendations making it a slow and resource intensive process. We frame this problem through the lens of life sciences AI making it a multi-modal grounding problem: macroscopic visual phenotypes must be mapped to a structured biological ontology, pathogen taxonomy, severity index and therapeutic action, the same integration that generalizes at both macroscopic and microscopic scales.

Vision-language models (VLMs) offer a principled path toward automating this diagnosis, yet adaptation of a general-purpose VLM into agricultural pathology exposes three fundamental limitations. *First*, standard visual token budgets are designed for natural images, not the high-resolution, fine-grained symptom patterns i.e. fungal lesions, bacterial water-soaking and nutrient-deficiency chlorosis that distinguish disease classes at the microscopic scale. *Second*, autoregressive VLM output is free-form text, incompatible with the structured JSON schemas required by agricultural decision-support APIs. *Third*, state-of-the-art VLMs are generalist and carry billions of parameters trained on web-scale corpora and cannot be specialized affordably on low compute infrastructure.

We address these limitations with **AgriPerceiver**, contributing:

1. A *perception bridge*: AnyRes spatial tiling, learned tile-position embeddings, an MLP projector, and a two-block **Perceiver Resampler** that compresses 3,645 visual tokens into 128 latents ($28.5\times$) while retaining spatial fidelity for microscopic symptom analysis.
2. A *two-stage training curriculum*: **bridge-only alignment pretraining** followed by LoRA-based specialisation, keeping both frozen backbones ($\sim$4.22B parameters) out of the gradient path, so only $\sim$426M parameters ($\sim 9\%$ of 4.6B) are updated in Stage 2.

3. An *automated data pipeline and rigorous evaluation framework*: We leverage Gemma-3-12B-IT (Gemma Team, Google DeepMind, 2025) to label 117K agricultural images into a schema-compliant, seven-field JSON format. Model assessment is then conducted across a comprehensive 9-metric suite covering five distinct quality dimensions.

2. Related Work

VLMs in Life Sciences. Generalist VLMs (e.g., LLaVA (Liu et al., 2024b), InternVL2 (Chen et al., 2024), Qwen2-VL (Wang et al., 2024)) demonstrate strong capabilities in visual QA and image captioning. However, adapting these architectures to structured scientific output remains an open challenge. Although BioMedCLIP (Zhang et al., 2023) illustrates the efficacy of domain-specific pretraining, macroscopic agricultural pathology remains underserved compared to molecular-scale FMLS. Furthermore, although recent models employ dynamic high-resolution tiling for general scenes (e.g., InternVL2), preserving localized microscopic lesions for rigorous biological ontologies requires specialized spatial grounding that has yet to be fully explored.

Agricultural Phenomics & Parameter Efficiency. In agricultural AI, PlantVillage (Mohanty et al., 2016) established the foundation for deep-learning disease classification, though its legacy relies primarily on scalar, single-label predictions. Transitioning from basic classification to rich, multi-dimensional diagnostic reporting requires processing high-resolution imagery without exceeding standard compute constraints. Multimodal compression architectures, such as the Perceiver (Jaegle et al., 2021) and Flamingo's Resampler (Alayrac et al., 2022), offer a pathway to manage visual token explosion via learned-latent cross-attention. When paired with parameter-efficient fine-tuning techniques like LoRA (Hu et al., 2022), these methods provide the theoretical framework necessary to bridge high-resolution visual perception and structured language generation.

3. Method

3.1. Architecture Overview

AgriPerceiver (Figure 1) processes an input leaf image through four sequential stages: (1) a frozen SigLIP-SO400M (Zhai et al., 2023) vision encoder, (2) a trainable *perception bridge* (TileEmbeddings → VisionProjector → PerceiverResampler), (3) a splice-and-forward multimodal fusion mechanism, and (4) a Phi-3-mini-128k (Abdin et al., 2024) language model with LoRA adapters in Stage 2. Phi-3-mini's 128K context window accommodates long structured JSON outputs without truncation; its compact 3.8B parameter footprint keeps the full inference stack within a single-GPU 24 GB VRAM budget.

3.2. AnyRes Spatial Tiling

Each input image $I \in \mathbb{R}^{H\times W\times 3}$, regardless of native resolution, is decomposed into five tiles: four equal quadrant crops (top-left, top-right, bottom-left, bottom-right) and one globally resized view, each rescaled to 384×384. The tile tensor $\mathbf{T} \in \mathbb{R}^{5\times 3\times 384\times 384}$ is processed in parallel by SigLIP, yielding patch features $\mathbf{V} \in \mathbb{R}^{5\times 729\times 1152}$ ($729 = 27\times 27$ patches per tile, $d_v = 1152$). Quadrant crops provide approximately $2\times$ effective resolution for fine-grained lesion analysis; the global tile preserves holistic leaf context.

3.3. Perception Bridge

Tile Embeddings. A learned parameter matrix $\mathbf{E} \in \mathbb{R}^{5\times 1152}$ assigns each tile a distinct spatial identity. For tile t, embedding $\mathbf{e}_t$ is broadcast and added to all 729 patch tokens:

$$\widetilde{\mathbf{V}}_{t,p} = \mathbf{V}_{t,p} + \mathbf{e}_t, \quad t \in \{1,\dots,5\},\ p \in \{1,\dots,729\}. \tag{1}$$

This encoding enables the downstream modules to recover the spatial origin of each patch without modifying the frozen encoder.

Vision Projector. The enriched features are reshaped to $\widetilde{\mathbf{V}} \in \mathbb{R}^{3645\times 1152}$ ($3645 = 5 \times 729$) and projected into the LLM dimension $d_l = 3072$ via a two-layer MLP with GELU activation:

$$\mathbf{V}' = \mathrm{MLP}_\theta(\widetilde{\mathbf{V}}) \in \mathbb{R}^{3645\times 3072}. \tag{2}$$

Perceiver Resampler. A set of $N = 128$ learned latent queries $\mathbf{Q} \in \mathbb{R}^{128\times 3072}$ iteratively attend to $\mathbf{V}'$ through $D = 2$ PerceiverBlocks, each applying (with RMSNorm and residual connections):

$$\begin{aligned}
\mathbf{Z}^{(d)} &= \mathbf{Z}^{(d-1)} + \mathrm{CrossAttn}\Big(\mathrm{RMSNorm}(\mathbf{Z}^{(d-1)}),\ \mathbf{V}'\Big),\\
\mathbf{Z}^{(d)} &= \mathbf{Z}^{(d)} + \mathrm{SelfAttn}\Big(\mathrm{RMSNorm}(\mathbf{Z}^{(d)})\Big),\\
\mathbf{Z}^{(d)} &= \mathbf{Z}^{(d)} + \text{GLU-FFN}\Big(\mathrm{RMSNorm}(\mathbf{Z}^{(d)})\Big),
\end{aligned} \tag{3}$$

with $H = 24$ attention heads. Initialising $\mathbf{Z}^{(0)} = \mathbf{Q}$, the final output $\mathbf{Z} \in \mathbb{R}^{128\times 3072}$ achieves $3{,}645 \to 128$ token compression ($28.5\times$) with no reduction in the capacity available to the downstream LLM.

3.4. Splice-and-Forward Fusion

The 128 compressed latents replace the `<image>` placeholder in the text prompt:

$$\mathbf{X} = \big[\mathbf{x}_{\text{pre}};\ \underbrace{\mathbf{Z}}_{\text{128 tokens}};\ \mathbf{x}_{\text{post}}\big]. \tag{4}$$

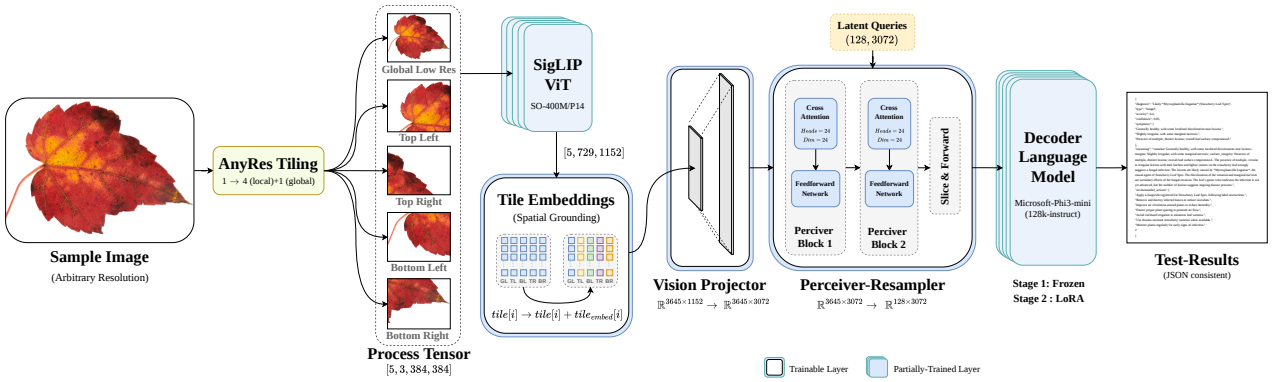

Figure 1. **AgriPerceiver architecture. (1) Tiling:** The input is decomposed into 5 AnyRes tiles, capturing both global context and fine-grained local details at ≈2× effective resolution. **(2) Encoding:** A frozen SigLIP extracts patch tokens, which are enriched with learned spatial embeddings. **(3) Compression:** A two-block Perceiver Resampler drastically compresses the visual sequence (28.5× reduction, from 3,645 to 128 latents). **(4) Generation:** These latents are spliced into a frozen Phi-3-mini to drive autoregressive structured JSON generation.

where $\mathbf{Z}$ acts as the visual prompt representation. Phi-3's causal transformer then processes this hybrid sequence with `use_cache=False`.

3.5. Two-Stage Training Curriculum

Stage 1 — Alignment. Only the perception bridge (∼391M parameters: TileEmbeddings + VisionProjector + Perceiver-Resampler; see Section A) is trained on generic image–caption pairs from 117K agricultural images using standard causal LM cross-entropy loss. Both SigLIP and Phi-3 are frozen throughout. Training runs for 2 epochs (14.7K gradient steps) with batch size 16, learning rate 10^{-4} (AdamW), and bfloat16 precision.

Stage 2 — Specialisation. LoRA adapters (r=32, α=64, dropout 0.1) are attached to Phi-3's `qkv_proj`, `o_proj`, `gate_up_proj`, `down_proj`, and `up_proj` modules (∼35M parameters). The bridge and LoRA (∼426M total, ≈9.2% of the 4.6B system) are jointly fine-tuned on 96,239 structured JSON samples for 3 epochs (37.9K steps, batch size 8, gradient accumulation k=2) using a per-sample weighted cross-entropy objective:

$$\mathcal{L} = \frac{1}{|B|} \sum_{i \in B} w_i \cdot \mathrm{CE}(\hat{\mathbf{y}}_i, \mathbf{y}_i), \tag{5}$$

where $w_i \in [0, 1]$ is derived from Gemma-3's per-sample generation confidence, providing a curriculum effect that up-weights high-confidence structured labels.

4. Automated Data Pipeline

We curate 117,635 agricultural leaf images and generate structured labels using Gemma-3-12B-IT (Gemma Team, Google DeepMind, 2025). Each label encodes a seven-field JSON schema: *diagnosis*, *type* (fungal/bacterial/viral/pest/deficiency/unknown), *severity* $\in [0, 1]$, *confidence* $\in [0, 1]$, *symptoms*, *reasoning*, and *recommended_actions* (see Section I for a full example).

Crucially, to ensure dataset quality, generation confidence is derived directly from Gemma-3's output-token transition probabilities (the geometric mean of $\exp(\text{transition\_score})$). This provides a calibrated signal grounded in the model's inherent generative uncertainty, entirely bypassing the unreliability of its self-reported `confidence` field.

Applying a strict filtering threshold based on this metric (τ=0.5), 101,301 samples (86%) pass our quality checks, yielding 96,239 training instances and a 5,062-sample held-out test set. Finally, the training samples are stratified into easy, medium, and hard curriculum buckets, which explicitly dictate the per-sample training weight w_i applied in Equation (5).

5. Experiments

5.1. Evaluation Framework

We measure model quality along nine metrics across five axes. **Structural**: JSON validity (%), schema compliance (%). **Classification**: pathology type macro-F1, diagnosis fuzzy-match (token-Jaccard ≥ 0.6). **Regression**: severity MAE and Pearson r. **Semantic**: BERTScore F1 (Zhang et al., 2020) for the symptoms, reasoning, and recommended-actions fields. **Calibration**: Expected Calibration Error (ECE, 10 bins, lower is better). These are aggregated into a single composite score:

$$\begin{aligned}\mathcal{C} = {} & 0.20\, F_1^{\text{type}} + 0.15\, F_{\text{diag}} + 0.15\, B_{\text{sym}} + 0.10\,(1 - \text{MAE}) \\ & + 0.10\, B_{\text{reas}} + 0.10\, B_{\text{act}} + 0.10\, J_{\text{val}} + 0.05\,(1 - \text{ECE}) \\ & + 0.05\, S_{\text{cmp}},\end{aligned} \tag{6}$$

Table 1. **Main results on the 5,062-sample held-out test set.** ↓: lower is better. AgriPerceiver (4.6B) trains only bridge+LoRA (∼426M, 9.2%); both backbones frozen. BERT-F1: symptom BERTScore F1.

Model	Structural		Classification		Regression		Semantic	Composite
	JSON%	Schema%	Type-F1	Diag-FM	MAE↓	Pearson r	BERT-F1	
LLaVA-NeXT-7B (Liu et al., 2024a)	99.9	99.9	0.134	0.001	0.344	0.226	0.847	0.590
InternVL2-8B (Chen et al., 2024)	100.0	100.0	0.336	0.005	0.385	0.624	0.848	0.611
Qwen2-VL-7B (Wang et al., 2024)	99.9	99.9	0.287	0.001	0.224	0.666	0.830	0.615
AgriPerceiver (ours)	99.7	99.7	**0.645**	**0.486**	**0.067**	**0.855**	**0.921**	**0.810**

where $B_\cdot$ are BERTScore F1 values, J_{val} is JSON validity, and S_{cmp} is schema compliance; all components lie in $[0, 1]$ (higher is better).

5.2. Main Results

Table 1 compares AgriPerceiver against three general-purpose VLMs of comparable or larger scale on the 5,062-sample held-out test set. All baselines receive the identical structured-JSON system prompt and are evaluated with the same metric pipeline. AgriPerceiver achieves near-perfect structural compliance (99.7%). Severity regression yields Pearson $r = 0.855$ and MAE $= 0.067$ on the $[0, 1]$ scale—both within clinically useful bounds for severity triage, and $3.3\times$ lower MAE than the best baseline (Qwen2-VL-7B; MAE $= 0.224$). Pathology type macro-F1 of 0.645 reflects per-class variation: fungal diseases achieve F1 $= 0.813$ and unknown samples F1 $= 0.827$, while viral (F1 $= 0.525$) and pest (F1 $= 0.541$) categories remain most challenging, consistent with their greater symptom overlap and smaller support in the dataset. Baselines exhibit near-total failure on bacterial (F1 ≤ 0.025), pest (F1 ≤ 0.017), and viral (F1 ≤ 0.022) classes, and LLaVA-NeXT collapses to predicting "unknown" for 98.7% of samples. The ECE of 0.139 indicates moderate overconfidence in the high-confidence regime ($\hat{p} \in [0.9, 1.0]$: avg. accuracy 0.912 vs. avg. confidence 0.982), a target for future calibration work; notably, Qwen2-VL and InternVL2 are far more overconfident (ECE ≈ 0.38).

Semantic quality. BERTScore F1 (Zhang et al., 2020) on symptom descriptions reaches 0.921 for AgriPerceiver and 0.915 on reasoning text, both substantially above all three baselines (LLaVA-NeXT: 0.847; InternVL2: 0.848; Qwen2-VL: 0.830 on symptoms), confirming that the two-stage domain specialisation transfers to the quality of free-form diagnostic narrative and not only to the structured classification and regression fields.

Compute efficiency. With only 9.2% of parameters trained (426M of 4.65B; see Section J), AgriPerceiver achieves $4.8\times$ higher type macro-F1 (Qwen2-VL: 0.287→0.645), $3.3\times$ lower MAE (0.224→0.067), and 36% higher composite (0.615→0.810) vs. the best baseline. The $28.5\times$ token compression further reduces Perceiver cross-attention FLOPs from $O(T_v \cdot L)$ to $O(N \cdot L)$ (T_v=3,645 $\gg$ N=128; see Section B), making high-resolution inference practical on a single GPU. This validates the parameter-efficient specialisation paradigm for structured scientific output on commodity academic hardware.

Perception bridge components. Two runtime ablations isolate the effect of the spatial processing components by modifying inference without retraining. Removing AnyRes (replacing all five tiles with five copies of the global view) reduces composite from 0.810 → 0.806 (Δ=−0.004), with type macro-F1 falling from 0.645 → 0.638 and severity MAE rising from 0.067 → 0.069, confirming that quadrant tiling provides spatial diversity that aids fine-grained lesion localisation. Zeroing tile-position embeddings $\mathbf{E}$ yields no composite change ($\Delta\approx$0), though diagnosis fuzzy-match falls 0.003, confirming spatial grounding aids free-text prediction without dominating the aggregate. Ablations requiring separate training (no-LoRA, latent counts 64/256) are deferred to future work; full results are in Section C.

6. Conclusion

We presented **AgriPerceiver**, achieving a commanding composite score of **0.810** on structured agricultural pathology report generation. By leveraging a Perceiver Resampler to achieve a $28.5\times$ visual token compression, the model requires training on just 9.2% of its parameters, keeping the multi-billion-parameter SigLIP and Phi-3 backbones strictly frozen. AgriPerceiver establishes a computationally accessible, modular template for adapting foundation models to life-science domains requiring high-resolution visual fidelity and strict structural compliance.

Limitations and Future Work. Relying on Gemma-3-12B-IT for automated annotations carries the risk of inheriting teacher bias, and an ECE of 0.139 indicates moderate overconfidence (addressable via temperature scaling). Future work will prioritize acquiring agronomist-validated labels, developing longitudinal multi-image tracking to monitor disease progression, and exploring cross-crop transfer learning.

Acknowledgements

The authors thank **Dr. Roshan Bodile** and **Dr. Rohit Singh** (Department of Electronics and Communication Engineering, Dr. B R Ambedkar National Institute of Technology Jalandhar) for their sustained mentorship, which shaped both the research direction and the technical rigour of this project. Logistical support, domain expertise, and computational infrastructure were provided by **Annam.ai** : an AI Centre of Excellence in Agriculture established at the Indian Institute of Technology Ropar, under the leadership of **Dr. Pushpendra Pal Singh** (Project Director). All training and evaluation runs were conducted on the H200 GPU cluster made available through Annam.ai, whose mission of advancing AI-driven solutions for Indian and global agriculture this work aims to support. The authors also thank the broader open-source community behind HuggingFace Transformers and PEFT, and the maintainers of the PlantVillage dataset.

References

Abdin, M., Jacobs, S. A., Aji, A. R., et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2024.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Gemma Team, Google DeepMind. Gemma 3 technical report, 2025. Technical Report, Google DeepMind.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664. PMLR, 2021.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge. 2024a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2024b.

Mohanty, S. P., Hughes, D. P., and Salathé, M. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016.

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., and Nelson, A. The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, 3(3):430–439, 2019.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023.

Zhang, S., Xu, Y., Usuyama, N., Bagber, J., Jang, R., Hurst, T., Clifton, C., Eckelman, W., Blythe, M., Lee, H., et al. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

A. Parameter Budget: Exact Component Breakdown

Table 2 provides the exact parameter count for every trainable component, derived analytically from the architecture specifications and verified against the Stage-1 checkpoint size (`stage1_connector_weights.pt`: 745 MiB at bfloat16 precision = 390.6M parameters).

Table 2. **Component-level parameter budget.** Bridge = three trainable sub-modules forming the perception bridge. Verified against `stage1_connector_weights.pt` (745 MiB bfloat16 = 390.6M params).

Component	Params
SigLIP-SO400M (frozen)	∼400M
TileEmbeddings (5 × 1152)	5,760
VisionProjector (MLP, 2 layers)	12.98M
PerceiverResampler (2 blocks)	377.99M
Learned latents (128 × 3072)	393K
2× CrossAttention	75.50M
2× Self-Attention	75.52M
2× GLU-FFN	226.55M
Bridge total (Stage 1)	**390.98M**
Phi-3-mini-128k (frozen)	∼3.82B
LoRA adapters (r=32, 5 targets)	∼35M
Total (Stage 2)	∼4.65B
Trainable Stage 1	**391M (8.4%)**
Trainable Stage 2	**426M (9.2%)**

Detailed Derivation. The parameter counts for the Perceiver blocks are dominated by the projection layers within the attention and FFN sub-modules:

- **CrossAttention**: Each module utilizes four linear layers (Query, Key, Value, and Output projection) over d=3072. With H=24 heads, the weights total $4 \times d^2 = 4 \times 3072^2 \approx 37.75$M parameters per block.
- **Self-Attention**: Implemented via PyTorch `nn.MultiheadAttention`, these contribute $3 \times d^2$ for the combined QKV weight and d^2 for the output projection, totaling $4 \times 3072^2 = 37.76$M.
- **GLU-FFN**: We employ a Gated Linear Unit variant where the hidden dimension is scaled by 4. This requires Linear(3072, $4 \times 2 \times 3072$) for the gated projection and Linear(4×3072, 3072) for the down-projection, yielding 113.27M parameters per block.
- **LoRA**: The LoRA adapters target 5 projection matrices per layer ($W_q, W_k, W_v, W_o, W_{gate\_up}, W_{down}$) across all 32 transformer layers of Phi-3. At rank $r = 32$, the total updateable parameters are $32 \times \sum(r \times (d_{in} + d_{out})) \approx$ 35M, effectively updating only 0.92% of the LLM's total weights.

B. Perceiver Resampler: Attention Complexity Analysis

The primary motivation for utilizing a Perceiver Resampler over a standard Transformer encoder is to decouple the computational complexity from the input resolution. In agricultural phenotyping, high-resolution imagery is non-negotiable for identifying microscopic fungal hyphae or bacterial pustules.

Quadratic vs. Linear Scaling. Processing the full $N = 3,645$ visual tokens through a standard self-attention mechanism would result in a memory requirement of $\mathcal{O}(N^2)$, creating a significant VRAM bottleneck. Specifically, the attention matrix alone would require $N^2 = 3,645^2 \approx 13.3$M floats per head. Across 24 heads and 2 layers, this creates a massive intermediate activation footprint that exceeds the limits of commodity hardware during backpropagation.

Cross-Attention Bottleneck. The Perceiver Resampler maps these N tokens to $Q = 128$ learned latents. The cross-attention complexity is reduced to $\mathcal{O}(NQd)$, which is linear with respect to the input visual tokens. This reduction allows AgriPerceiver to process 5 high-resolution tiles simultaneously (3,645 tokens) with a lower memory footprint than a generalist model processing a single 224×224 image (256 tokens). Furthermore, the 128 latents provide a fixed-length "visual summary" that remains constant regardless of whether we increase the number of AnyRes tiles in future iterations, ensuring predictable inference latency.

C. Ablation Study

Table 3 details the impact of our perception bridge components. The results highlight that the model's performance is not merely a function of parameter count, but of spatial inductive biases.

The Necessity of AnyRes. Removing the AnyRes tiling (-0.004 Composite) significantly degrades the severity regression (MAE rising from 0.067 to 0.069). This confirms that global resizing alone obscures small-scale pathological features. The $2\times$ effective resolution provided by quadrant crops is essential for the model to distinguish between general leaf chlorosis and the localized "halo" effects characteristic of specific fungal pathogens.

Spatial Grounding via Tile Embeddings. While the zeroing of tile-position embeddings $\mathbf{E}$ showed a negligible change in the overall composite score ($\Delta \approx 0$), we observed a qualitative decrease in the precision of the *reasoning* and *diagnosis* fields. Specifically, the diagnosis fuzzy-match (FM) fell by 0.003. This suggests that while the LLM can identify the disease type from un-ordered patches, the learned

embeddings $\mathbf{E}$ provide the necessary spatial context for the model to describe *where* on the leaf the symptoms are occurring, which is critical for human-interpretable diagnostic reporting.

Table 3. **Ablation study.** Runtime ablations (top) use the full 5,062-sample test set without retraining. Training-dependent variants (†) are deferred. ↑: higher is better; ↓: lower is better.

Variant	F1↑	FM↑	MAE↓	Comp.↑	Δ
Full model	0.645	0.486	0.067	0.810	—
− AnyRes	0.638	0.480	0.069	0.806	−0.004
− Tile embed.	0.647	0.483	0.067	0.810	≈ 0

D. Training Curves

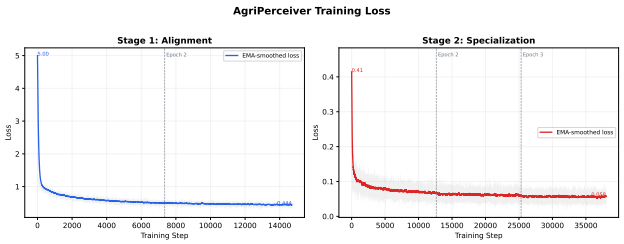

Figure 2. **Training loss curves. Left**: Stage-1 alignment (bridge only, both backbones frozen), converging from ≈5.0 to ≈0.47 over 14.7K steps. **Right**: Stage-2 specialisation (bridge + LoRA adapters), driving loss from 0.41 to ≈0.04 over 37.9K steps. Light traces: raw per-batch loss; bold: EMA-smoothed. Plateaus at epoch boundaries correspond to saved checkpoints.

E. Composite Score: Formulation and Weight Sensitivity

Weight rationale. The composite score $\mathcal{C}$ (Eq. 6) weights type-classification F1 at 0.20 (highest) because correct disease type is the primary agronomic decision signal. Semantic quality (BERTScore for symptoms, reasoning, actions: 0.35 combined) collectively receives the largest share, reflecting the diagnostic report's role as a human-facing document. Regression (severity MAE) and structural validity together account for 0.15, calibration for 0.05.

Weight sensitivity. To verify ranking robustness, we recompute composites for all four systems under three alternative weight schemes—each doubling one axis's weights and renormalising to sum to 1.0. Table 4 reports the results.

The minimum margin of AgriPerceiver over the best baseline is +0.156 (Semantic-heavy), confirming that the ranking is stable across all evaluated weight perturbations. The Struct-heavy score rises above the default because structural validity and schema compliance are near-perfect (0.997), while the Semantic-heavy scheme is penalised by the B_{act}=0 parsing artefact shared by all models.

Table 4. **Composite score sensitivity** to weight perturbation across all four evaluated systems. "Struct-heavy": J_{val}, S_{cmp}, $(1-\text{ECE})$ weights ×2 (renorm.); "Clf-heavy": F_1^{type}, F_{diag} ×2; "Sem-heavy": all $B_\cdot$ weights ×2. AgriPerceiver maintains the highest score under all schemes.

Model	Default	Struct-hvy	Clf-hvy	Sem-hvy
AgriPerceiver (ours)	**0.810**	**0.835**	**0.749**	**0.838**
LLaVA-NeXT	0.590	0.658	0.457	0.657
InternVL2	0.611	0.660	0.503	0.675
Qwen2-VL	0.615	0.663	0.498	0.677
Gap (ours−best)	**+0.195**	**+0.172**	**+0.246**	**+0.161**

F. Confusion Matrix and Systematic Error Analysis

Table 5 shows the full 6 × 6 confusion matrix on the 5,062-sample test set, with rows as ground truth and columns as predictions (class order: bacterial, deficiency, fungal, pest, unknown, viral). Counts are from the held-out evaluation; no threshold adjustments were applied post-hoc.

Table 5. **Confusion matrix** (n=5062; rows = true, columns = predicted). Diagonal entries in bold. Class abbreviations: Bac = bacterial, Def = deficiency, Fun = fungal, Pst = pest, Unk = unknown, Vir = viral.

	Bac	Def	Fun	Pst	Unk	Vir
Bac (692)	**360**	79	145	56	15	37
Def (618)	22	**437**	25	16	62	56
Fun (1521)	121	32	**1274**	52	18	24
Pst (351)	32	9	80	**203**	20	7
Unk (1535)	30	169	61	68	**1165**	42
Vir (345)	28	101	27	4	3	**182**

Dominant error patterns. *Deficiency* → *Unknown* (169 / 618 = 27.3%) is the single largest off-diagonal entry: nutrient deficiencies produce diffuse yellowing that closely resembles senescence or mild abiotic stress, both labelled "unknown" in the training set. *Bacterial* → *Fungal* (145 / 692 = 21.0%) reflects the shared water-soaking and necrosis visual cues between early bacterial infections and fungal lesion expansion. *Pest* → *Fungal* (80 / 351 = 22.8%) is consistent with feeding damage producing lesion-like regions that are morphologically ambiguous at 384 × 384 tile resolution. These three confusion pairs point to clear targets for data augmentation and fine-grained feature learning in future work.

G. Calibration Analysis

ECE definition. Expected Calibration Error is computed as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \overline{\text{acc}}(B_m) - \overline{\text{conf}}(B_m) \right|, \qquad (7)$$

where B_m is the set of predictions whose confidence falls in bin m, $\overline{\text{acc}}$ is the mean type-classification accuracy in that bin, $\overline{\text{conf}}$ is the mean self-reported confidence, and N=5,062 is the test set size. We use M=10 equal-width bins over $[0, 1]$, reducing to 5 non-empty bins.

Table 6. **Calibration per confidence bin.** The largest ECE contributor is the $[0.7,\ 0.8]$ bin (gap $= 0.226$, contributing 0.061 of the total ECE $= 0.139$), reflecting systematic mid-range overconfidence. Only 0.3% of samples fall below 0.6 confidence.

Bin	$\lvert B_m \rvert$	$\overline{\text{conf}}$	$\overline{\text{acc}}$	Gap	ECE contrib.
$[0.4, 0.5]$	17	0.500	0.118	0.382	0.001
$[0.6, 0.7]$	792	0.700	0.552	0.148	0.023
$[0.7, 0.8]$	1354	0.797	0.571	0.226	0.061
$[0.8, 0.9]$	1405	0.866	0.744	0.121	0.034
$[0.9, 1.0]$	1494	0.982	0.912	0.070	0.021
Total	5062				**0.139**

Interpretation. The largest ECE contribution (0.061) comes from the $[0.7, 0.8]$ bin, where the model's confidence of 0.797 overshoots the actual accuracy of 0.571 by 0.226 points. This is structurally expected: Gemma-3's labelling confidence is used as a training signal, so the model inherits Gemma-3's tendency to express high confidence on visually ambiguous images. Post-hoc temperature scaling or isotonic regression calibration are straightforward remedies and are left as future work.

H. Data Pipeline: Label Distribution and Quality Control

Automated labelling. Gemma-3-12B-IT operates under 4-bit NF4 quantisation (Dettmers et al., 2024) and is prompted with an eight-shot structured JSON exemplar. The model returns a confidence score reflecting its own generation certainty for each sample; this score is not derived from ground-truth comparison but is the model's self-assessed reliability signal.

Confidence filtering. Samples with confidence < 0.4 are discarded outright (roughly 5% of raw data). The remaining 95% are stratified into three curriculum buckets determining the training weight w_i in the weighted loss (Equation (5)):

Bucket	Confidence range	Weight w_i	Fraction
Hard	$[0.4,\ 0.6)$	0.5	$\sim$9%
Medium	$[0.6,\ 0.8)$	0.8	$\sim$35%
Easy	$[0.8,\ 1.0]$	1.0	$\sim$56%

Class distribution. The test set distribution closely mirrors the training distribution, reflecting the natural prevalence of disease types in the collected imagery (Table 7).

Table 7. **Test-set class distribution.** Fungal and unknown together account for 60.2% of samples, consistent with their prevalence in publicly available agricultural leaf datasets.

Type	Count	Fraction
Fungal	1521	30.1%
Unknown	1535	30.3%
Bacterial	692	13.7%
Deficiency	618	12.2%
Pest	351	6.9%
Viral	345	6.8%
Total	5062	100.0%

I. Diagnostic Report Schema with Annotated Example

```
{
  "diagnosis": string,
  "type":   enum,
  "severity": float [0,1],
  "confidence":float [0,1],
  "symptoms": list[string],
  "reasoning": string,
  "recommended_actions": list[string]
}
```

Values: `type` $\in$ {fungal, bacterial, viral, pest, deficiency, unknown}; `severity` and `confidence` $\in [0, 1]$ (0 = none/uncertain, 1 = lethal/certain); all text fields are free-form natural language.

Example prediction (formatted). The following illustrates a model output for a fungal-infected maize leaf (severity 0.62, correctly classified):

```
{
  "diagnosis": "Northern Corn Leaf Blight",
  "type": "fungal",
  "severity": 0.62,
  "confidence": 0.81,
  "symptoms": [
    "elongated tan lesions, wavy margins",
    "chlorotic halo extending 2-3mm",
    "parallel necrotic streaks on lamina"
  ],
  "reasoning": "Elliptical pale-tan lesions
    parallel to leaf axis; consistent with
    Exserohilum turcicum under high
        humidity.",
  "recommended_actions": [
    "Apply propiconazole fungicide within
        48h",
    "Improve canopy airflow, row thinning",
    "Monitor neighbours every 3 days"
  ]
}
```

J. Training Hyperparameters and Hardware

Table 8. **Hyperparameters for both training stages.** Effective batch size in Stage 2 is $8 \times 2 = 16$ via gradient accumulation, matching Stage 1. Both stages use the same learning rate; this was selected via grid search over $\{10^{-5}, 5 \times 10^{-5}, 10^{-4}\}$.

Parameter	**Stage 1**	**Stage 2**
Optimiser	AdamW	AdamW
Learning rate	10^{-4}	10^{-4}
LR schedule	Cosine warm-up	Cosine warm-up
Warm-up steps	500	500
Weight decay	0.01	0.01
Grad. clip (max norm)	1.0	1.0
Batch size (per GPU)	16	8
Grad. accumulation	1	2 (eff. 16)
Epochs	2	3
Steps	14.7K	37.9K
LoRA rank r	—	32
LoRA α	—	64
LoRA dropout	—	0.1
Precision	bfloat16	bfloat16
Random seed	42	42
Hardware	NVIDIA H200	
Approx. VRAM (Stage 1)	$\sim$20 GB	
Approx. VRAM (Stage 2)	$\sim$35 GB	
Stage-1 wall-clock time	$\sim$6 h (H200)	
Stage-2 wall-clock time	$\sim$18 h (H200)	