Fine-Tuning of BioCLIP 2.5 with Taxonomic Heads for Multi-Species Plant Identification

Arjun Raj[1,†], Manindra de Mel[1,*,†], Razeen Wasif[1,†] and William Brake[1,†]

[1]The Australian National University, Canberra, ACT 2601, Australia

Abstract

We describe ANU's submission to PlantCLEF 2026, where the task is to predict every species present in a vegetation-quadrat image. The central challenge is a domain shift where training images show single plants while test images show dense multi-species plots; this is further compounded by a long-tailed label distribution. Our system is BioCLIP 2.5 ViT-H/14 where only the last four transformer blocks plus the final layer norm and projection are unfrozen, while the lower twenty-eight blocks stay frozen to preserve the Tree-of-Life prior. We train on an enlarged manifest that combines PlantCLEF 2024 with a research-grade iNaturalist pull (2.65 million single-plant images, 7,806 species) with auxiliary genus- and family-level cross-entropy heads that regularise the species head without adding inference cost. At inference each quadrat is split into a 4×4 grid of sixteen tiles, each tile is forwarded through the encoder at both 224 and 336 pixels, the per-tile softmax distributions are averaged across tiles and across resolutions, class-prior logit adjustment is applied, and an adaptive probability threshold selects the emitted species set. Our best public Macro F1 is **0.41826**; the same configuration scores **0.40283** private (the strongest private result of the project, 0.40600, comes from a logit-adjusted single-resolution variant). Our team, **ARM Wision**, placed 7th on the official PlantCLEF 2026 leaderboard. Project website: https://arm-wision.github.io/; code: https://github.com/arm-wision/PlantCLEF2026.

Keywords

plant identification, multi-label classification, partial fine-tuning, BioCLIP, taxonomic auxiliary loss, vegetation plot analysis, PlantCLEF 2026

1. Introduction

Monitoring plant biodiversity at scale is essential for understanding ecosystem health, tracking the effects of climate change, and informing conservation policy. Traditionally, vegetation surveys require expert botanists to visit field sites, place standardised quadrats (typically 50×50 cm sampling frames), and manually identify every plant species visible within the frame. This process is labour-intensive, difficult to scale, and constrained by the availability of trained taxonomists [1].

The PlantCLEF 2026 challenge [2, 1] seeks to automate this task by framing it as a multi-label classification problem: given a high-resolution photograph of a vegetation quadrat, predict the set of all plant species present. The challenge draws from a taxonomy of approximately 7,800 candidate species, and submissions are evaluated using a macro-averaged F1 score computed per quadrat image and then averaged across transects:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{j=1}^{T_i}\mathrm{F1}^{j}\right),$$

where N is the number of transects, T_i is the number of quadrats in transect i, and $\mathrm{F1}^{j} = \frac{2\cdot P_j \cdot R_j}{P_j + R_j}$ is the per-image F1 score, with P_j and R_j the precision and recall over the predicted species set for quadrat j. Averaging first within transects then across them prevents high-density sampling sites from dominating the final score.

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

u7526852@anu.edu.au (A. Raj); u7543337@anu.edu.au (M. d. Mel); u7619822@anu.edu.au (R. Wasif); u7480294@anu.edu.au (W. Brake)

A central difficulty of this challenge is the *domain shift* between the available training data and the target test data. The training set consists of over one million single-plant images from the PlantCLEF 2024 corpus, each depicting an individual specimen with a single species label (Figure 1). In contrast, the test images are dense, complex vegetation scenes where multiple species overlap, occlude one another, and appear at varying scales and growth stages (Figure 2). Models trained exclusively on isolated specimens tend to perform poorly when confronted with these multi-species scenes, as they have never encountered the visual complexity of real-world vegetation during training.

Figure 1: Six representative single-plant training images from the PlantCLEF 2024 corpus. Each image carries a single species label and was contributed through the Pl@ntNet citizen-science pipeline. Shooting style is intentionally varied (close-up macro, half-body, full-plant in habitat) and lighting / background are uncontrolled.

A second major challenge is the *long-tail distribution* of the training data. While common species may have thousands of training images, rare species, which are often the most ecologically important, can have fewer than ten. Standard cross-entropy training causes models to overwhelmingly predict common classes, yielding high micro-accuracy but poor macro-F1.

In this paper, we present the system submitted by the Australian National University (ANU) team, competing under the team name **ARM Wision**, under which our results appear on the official PlantCLEF 2026 leaderboard. Our approach centres on three key design decisions:

1. **Partial-unfreeze BioCLIP 2.5 with taxonomic regularisation.** BioCLIP [3] is a contrastive vision-language model whose pretraining on TreeOfLife-10M aligns its feature space with the Linnaean taxonomy, and BioCLIP 2 [4] scales this recipe to TreeOfLife-200M. Rather than fine-tuning the full backbone (which we show collapses this prior), we unfreeze only the last four transformer blocks of the BioCLIP 2.5 ViT-H/14 checkpoint [4] and attach auxiliary cross-entropy heads at coarser taxonomic levels (genus and family in the final design; an earlier multi-task fine-tune additionally used order and class). The auxiliary losses carry small fixed weights, with coarser levels weighted progressively less, so that they primarily regularise the encoder rather than drive it.
2. **Long-tail data rebalancing.** The PlantCLEF 2024 single-plant corpus is heavily skewed: a small number of species exceed one thousand training images while hundreds carry fewer than ten. We attack this primarily by enlarging the training manifest to 2.65 million images with pre-filled genus and family labels for every row. A follow-up experiment additionally capped each species at 500 images to flatten the head of the distribution, but this hurt rather than helped (0.40041 vs. 0.41826 public Macro F1), evidence that on this benchmark the marginal value of additional head-class examples is positive when label quality is high, and that flattening the distribution discards information rather than rebalancing it.
3. **Tile-level inference with adaptive selection.** At inference we partition each quadrat into a non-overlapping 4×4 grid of sixteen tiles, forward each tile through the encoder at both 224 and 336 pixels, and aggregate the per-tile softmax distributions by per-species mean across both the sixteen tiles and the two resolutions. Class-prior logit adjustment (τ=0.25) is then applied, which down-weights common species to counteract the long-tail bias. Final selection is adaptive: we emit every species whose post-adjusted probability exceeds a fixed threshold (T=0.03), with a floor of two and a ceiling of ten species per quadrat. The dual-resolution ensemble (224 and 336 pixels) lifts the public score by +0.0066 over the single-resolution baseline; the adaptive threshold outperforms the best fixed top-k (k=2) by about 0.007 Macro F1, and by larger margins for other k.

Figure 2: Six representative quadrats from the PlantCLEF 2026 test set, illustrating the breadth of within-set domain variation. Filenames (left to right): `CBN-PdlC-A1-20130807`, `CBN-Pla-A1-20130808`, `GUARDEN-CBNMed-10-1-16-34-20240429`, `LISAH-JAS-0-1-20220505`, `OPTMix-0108-P1-312-20231006`, `RNNB-1-1-20230512`. Each filename encodes the transect site and observation date. The quadrats range from dense mid-summer canopies (CBN-PdlC, RNNB) to dry shrub-mat communities (LISAH, GUARDEN) and litter-covered shoulder-season frames (OPTMix). Shooting angle, sampling-frame style (wood, tape, painted line), and illumination all vary across the corpus, compounding the single-plant → multi-species domain shift seen in Figure 1.

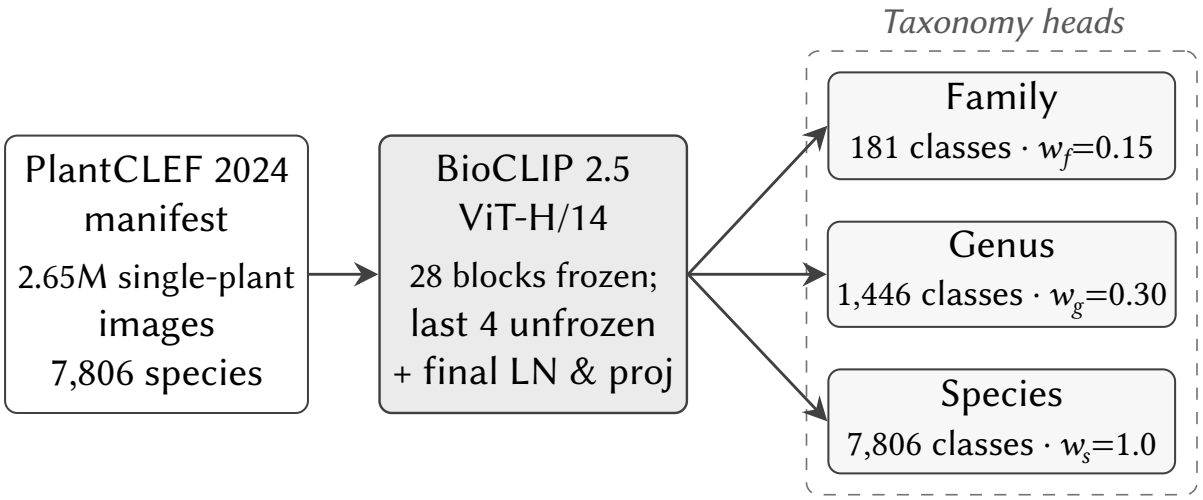

Figure 3: Training pipeline for the ANU PlantCLEF 2026 submission. The PlantCLEF 2024 single-plant manifest (2.65M images across 7,806 species) is fed into a partially fine-tuned BioCLIP 2.5 ViT-H/14 in which the lower 28 transformer blocks are frozen to preserve the Tree-of-Life prior and only the last 4 blocks plus the final layer norm and projection are unfrozen. Auxiliary cross-entropy heads at the family, genus, and species levels are attached.

The remainder of this paper is organised as follows. Section 2 reviews related work on plant identification, foundation model adaptation, and multi-label classification. Section 3 defines the final system top-down, from data and label construction through to the final inference pipeline. Section 4 sets out the evaluation protocol, the baseline ladder, and the experiments that justify each component. Section 5 analyses why the system behaves as it does: the validation/leaderboard decoupling, the head representation geometry, the within-backbone saturation, and the long-tail behaviour. Section 6 discusses limitations and future work, and Section 7 concludes.

2. Related Work

2.1. Automated Plant Identification

Large-scale plant identification has been driven by the LifeCLEF series of evaluation campaigns [1], which have progressively increased in scale and difficulty. Early systems relied on hand-crafted features such as leaf shape descriptors and colour histograms [5]. The advent of deep convolutional networks brought substantial improvements, with Inception and ResNet architectures achieving competitive results on single-specimen classification [6]. More recent work has shifted toward Vision Transformers (ViTs), which excel at capturing long-range spatial dependencies within images [7].

The PlantCLEF 2024 and 2025 editions introduced the vegetation plot setting, highlighting the domain gap between specimen-level training data and plot-level test data [8]. Top-performing systems in those editions employed strategies such as tiled inference (SAHI-style slicing), test-time augmentation, and ensemble methods to bridge this gap.

2.2. Foundation Models for Biodiversity

BioCLIP [3] is a contrastive vision-language model trained on the BIOSCAN and TreeOfLife-10M datasets, which encompass millions of organism images annotated with hierarchical taxonomic labels. Unlike general-purpose CLIP models, BioCLIP aligns visual features with the Linnaean taxonomy, making its feature space inherently suited for biological classification tasks. Studies have shown that BioCLIP features outperform ImageNet-pretrained features for species identification, particularly for morphologically similar species [3]. BioCLIP 2 [4] scales this hierarchical contrastive recipe to the 214-million-image TreeOfLife-200M dataset and reports emergent structure in the embedding space (inter-species features aligning with ecological meaning while intra-species variation is preserved in orthogonal subspaces). The backbone we fine-tune is the BioCLIP 2.5 ViT-H/14 checkpoint from this line of models [4].

DINOv3 [9] is a self-supervised ViT trained with a student-teacher distillation objective on a large curated dataset. Its features are highly transferable across domains and exhibit strong spatial correspondence, making them effective for fine-grained recognition tasks. We use DINOv3 features as one of three frozen feature extractors in our triple-backbone late-fusion classifier (Section 4), where the three backbones' features are concatenated rather than combined through any architectural change to the backbones themselves.

ConvNeXt-V2 [10] modernises the convolutional neural network paradigm by incorporating design elements from Transformers (e.g., larger kernel sizes, LayerNorm, GELU activations) alongside a Fully Convolutional Masked Autoencoder (FCMAE) pretraining strategy. ConvNeXt-V2 retains the inductive biases of CNNs (translation equivariance and locality) while achieving performance competitive with ViTs.

2.3. Partial Fine-Tuning of Large Pretrained Models

Fine-tuning all parameters of large pretrained models is often impractical: it requires substantial memory, risks catastrophic forgetting of pretraining knowledge, and tends to overfit when the downstream training set is noisy or long-tailed. A widely used alternative is to freeze the bulk of the backbone and adapt only the top few transformer blocks, which preserves the lower-layer features that capture domain-general structure while allowing the upper layers to specialise. For taxonomically aligned foundation models such as BioCLIP, this trade-off is particularly favourable, since the pretraining prior itself is what makes the model valuable on biological tasks.

2.4. Multi-Label Classification and Long-Tail Learning

Multi-label classification requires predicting a set of labels per input, typically using sigmoid activations and binary cross-entropy. Asymmetric Loss (ASL) [11] extends focal loss [12] to the multi-label

setting by applying separate focusing parameters γ_+ and γ_- for positive and negative samples. By setting $\gamma_- > \gamma_+$, ASL aggressively down-weights easy negatives, which dominate in settings with thousands of classes.

Logit adjustment [13] provides a complementary approach by shifting logits based on class prior probabilities, effectively requiring the model to produce larger margins for common classes and smaller margins for rare ones.

3. Method

We describe the final system top-down, as it would be reproduced, rather than in the order we discovered it; the chronological development trace of every direction we explored is given in Appendix A. The system has three parts. First, a training corpus of single-plant images carrying taxonomic labels (§3.1). Second, a partial fine-tune of BioCLIP 2.5 ViT-H/14 that adapts only the last four transformer blocks and attaches auxiliary taxonomy heads, preserving the Tree-of-Life prior that makes the backbone valuable on biological data (§3.2–§3.3). Third, a test-time pipeline that tiles each quadrat, pools the per-tile votes by softmax mean, calibrates with class-prior logit adjustment, and selects an adaptive number of species (§3.4–§3.6). We close with a single canonical statement of the final configuration (§3.7). Throughout this section we define *what* the system is and defer all leaderboard numbers to Section 4.

3.1. Data and Label Construction

The training data are single-plant images from the PlantCLEF 2024 corpus, each contributed through the Pl@ntNet citizen-science pipeline and carrying one species label together with its Linnaean ancestry (genus, family, and, where available, order and class). The distribution is heavily long-tailed: a small number of common species exceed one thousand images while hundreds of rare species carry only a handful, so vanilla shuffled training exposes the network to roughly two orders of magnitude more head-class than tail-class examples per epoch.

Our final model trains on an enlarged and cleaned manifest. The original PlantCLEF 2024 metadata (merged with a GBIF taxonomy lookup) supplies the first fine-tune; a subsequent data rebuild (experiment i001) re-downloaded the corpus, supplemented it with a research-grade iNaturalist pull, deduplicated, and pre-filled genus and family labels for every row, yielding roughly 2.65 million images across 7,806 species (Table 1). Our final model (experiment i002) trains on this larger manifest *without* any per-species cap. We also explored an explicitly rebalanced variant (experiment i003) that augments the tail with extra validated images and caps every species at five hundred images; its leaderboard outcome, which did not improve on the uncapped manifest, is analysed in Section 5.4. We therefore treat "more clean data, no cap" as a deliberate design choice for the final system rather than as a controlled study of rebalancing.

3.2. BioCLIP 2.5 Fine-tuning

BioCLIP [3] is a contrastive vision-language model whose pretraining on TreeOfLife-10M aligns its feature space with the Linnaean taxonomy, which is precisely the prior that makes it valuable for biological classification; BioCLIP 2 [4] scales this recipe to the 214-million-image TreeOfLife-200M dataset. We adapt the BioCLIP 2.5 ViT-H/14 image encoder [4] (hf-hub:imageomics/bioclip-2.5-vith14) by supervised fine-tuning, but only partially: rather than updating the full backbone, we freeze the lower twenty-eight transformer blocks and unfreeze only the last four blocks together with the final layer norm and projection layer. This narrow trainable surface is enough to adapt the encoder to the single-plant distribution while preserving the broader biological prior, and it is more resilient to optimiser noise on a long-tailed training set than full fine-tuning. The number of unfrozen blocks is itself a sharp design variable; we report the ablation that locates its optimum in Section 4.

Source / Split	Images	Observations	Species	Genera
PlantCLEF 2024 (Pl@ntNet)	1,408,033	1,151,904	7,806	1,446
iNaturalist (research-grade)	1,245,748	1,245,085	5,037	1,254
i001 manifest (union)	**2,653,781**	—	**7,806**	**1,446**
Train (stratified 90%)	2,391,457	1,121,447	7,806	1,446
Val (stratified 10%)	262,324	123,768	7,309	1,429

Table 1
Composition of the i001 training manifest used by the i002 model. The PlantCLEF 2024 row reproduces the dataset overview from [8]; the iNaturalist row reflects the research-grade pull we de-duplicated and image-verified before adding to the manifest. All iNaturalist species are a subset of the PlantCLEF 2024 taxonomy, so the union species and genus counts match the PlantCLEF 2024 row; iNaturalist therefore covers 5,037 of the 7,806 classes (the most populous head) rather than extending the long tail. The Train / Val rows are the deterministic stratified split applied at training time (10% per species, seed 42, species with fewer than five images held entirely in Train); the Val row drops to 7,309 species because 497 rare species fall below the five-image threshold and remain Train-only. Observations are reported only on the union and split rows where the field is well defined for iNaturalist sources, and they are omitted on the union total because the field is undefined for PlantCLEF 2024 rows in our merged manifest.

Training is staged in two phases. An initial ten epochs of head-only training on the enlarged manifest with the backbone fully frozen initialises the per-head MLPs of §3.3; a further ten epochs with the last four transformer blocks unfrozen then produce the final i002 checkpoint, resuming the model weights only. Both stages train in bfloat16 across two RTX 5090 GPUs under cross entropy with linear warmup and cosine decay; the complete hyperparameter specification (learning rates, batch sizes, gradient accumulation, augmentations, image transforms, and the validation split) is given in Appendix B, Table 12.

3.3. Taxonomy-aware Heads

On top of the encoder we attach a small head that predicts species jointly with coarser taxonomic levels, so the auxiliary taxonomy losses act as a structural regulariser. Each head is built from the same template (layer normalisation of the 1,024-dimensional CLS embedding, a linear projection to a 1,024-dimensional hidden representation, a GELU non-linearity, and dropout at rate 0.2), but the two fine-tunes we report wire it differently.

Shared trunk (experiment 010). Our first strong fine-tune uses a single *shared* multi-layer perceptron whose one hidden vector is read by every taxonomic head, so the auxiliary losses regularise the exact representation the species head consumes. It exposes five heads, with output dimensions 7,806 (species), 1,446 (genus), 181 (family), 61 (order), and 6 (class).

Per-head MLPs (experiment i002). Experiment i002 instead gives each head its own *independent* multi-layer perceptron of the same structure, decoupling the per-task gradients above the backbone. Because the redownloaded i001 manifest carries species, genus, and family columns but no order or class columns, it retains three per-head branches and drops the two coarsest levels.

The training objective is a weighted joint cross entropy

$$\mathscr{L} \;=\; \mathscr{L}_{\text{species}} \;+\sum_{l\in L} w_l\,\mathscr{L}_l, \tag{1}$$

where the species term carries an implicit weight of one, $L = \{g, f, o, c\}$ for the shared-trunk fine-tune and $L = \{g, f\}$ for the i002 per-head model, and the auxiliary weights are w_g=0.30, w_f=0.15, w_o=0.05, w_c=0.02. These are fixed, hand-set constants rather than learned or entropy-derived weights; coarser levels are deliberately down-weighted so that, when present, the order and class heads act primarily as regularisation rather than driving the encoder. The i002 per-head model uses only w_g=0.30 and w_f=0.15. Missing taxonomy labels are encoded as -1 and masked out of the auxiliary loss.

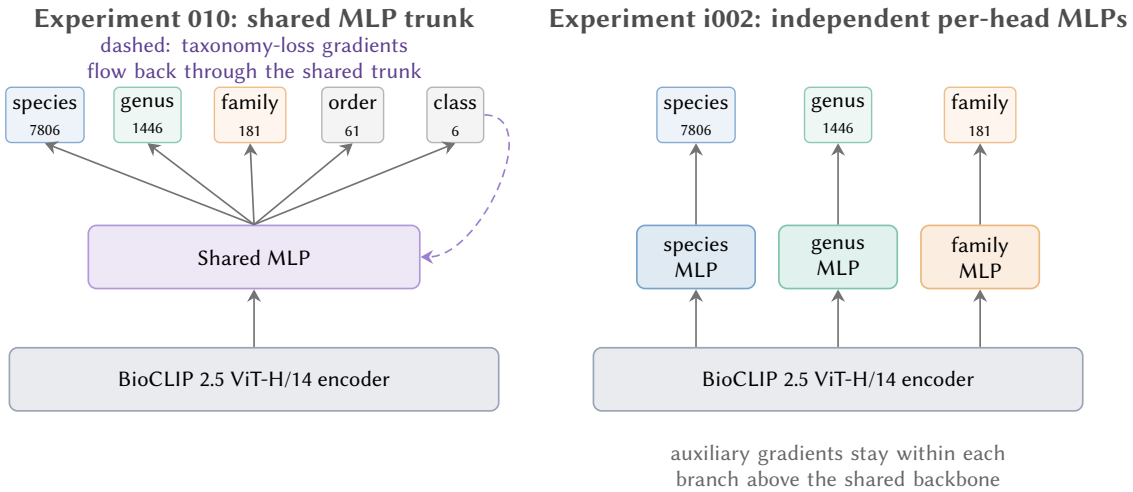

Figure 4: Head designs of the two supervised fine tunes. Experiment 010 (left) uses a single shared MLP (layer normalisation, a linear projection of the 1,024-d embedding, GELU, and dropout) whose hidden vector feeds five linear heads (species/genus/family/order/class); the auxiliary taxonomy losses back propagate through the shared trunk. Experiment i002 (right) uses three independent per head MLPs (species/genus/family), so the auxiliary gradients do not pass through the species MLP. Both share the partially unfrozen BioCLIP 2.5 backbone (last blocks plus the final layer norm and projection). The diagram shows the design difference only: the two experiments are not a controlled ablation, since they also differ in training data, schedule, unfrozen blocks, and head count.

We stress that the shared-trunk and per-head designs differ in several other respects at the same time (training data, schedule, number of unfrozen blocks, and head count), so we do *not* run a controlled comparison that varies only the head wiring, and we make no claim that the per-head MLP, or the taxonomy supervision, caused any performance gain. We report the head wiring as a deliberate design choice and analyse the resulting feature spaces descriptively in Section 5.2.

3.4. Tiled Inference

The classifier is trained on single-plant images, but each test quadrat is a dense, multi-species scene in which plants overlap, occlude one another, and appear at varying scales. Passing a whole quadrat through an encoder trained on isolated specimens therefore mixes many species into one embedding. We instead partition each quadrat into a non-overlapping 4×4 grid of sixteen tiles, each $1/4$ of the quadrat in width and height (and $1/16$ of its area), so that most tiles contain only one or a few species, closer to the training distribution. Each tile is forwarded through the encoder at two resolutions, the encoder's native 224 pixels and 336 pixels (the ViT-H/14 backbone tolerates the 1.5× stretch via bicubic positional-embedding interpolation), giving a per-tile species distribution that the next stage pools.

The 4×4 grid and the $\{224, 336\}$ resolutions were chosen as an initial compromise rather than tuned exhaustively. A coarse grid does risk down-sampling small plants: each tile covers a quarter of the quadrat in both dimensions, and on a typical ~2000-pixel-wide quadrat a single ~500-pixel tile is then resized to 224 or 336 pixels, so a small seedling occupying a few tens of pixels can be blurred away. We chose the configuration against three competing pressures. First, a denser grid (e.g. 6×6 or 8×8) multiplies the per-quadrat forward count quadratically ($16 \rightarrow 36 \rightarrow 64$ tiles, each run at two resolutions), which our submission-time compute budget did not allow at scale. Second, finer tiles increasingly slice individual plants across tile boundaries, fragmenting a single specimen's evidence across neighbouring tiles and working against the very single-plant framing the tiling is meant to recover. Third, the encoder is trained at 224 pixels, so the 224/336 pair keeps the inputs close to the training resolution while the 336-pixel pass recovers some fine detail (its empirical lift is reported in §4.6). We did *not* run a controlled sweep over grid density or tile size on the final checkpoint, again for want of submission budget and time, so we present 4×4 at $\{224, 336\}$ as a deliberate but unverified operating point rather

than a demonstrated optimum; a systematic grid-density and tile-size ablation is an explicit item of future work (§6).

3.5. Tile Aggregation and Selection

The sixteen per-tile predictions must be pooled into one image-level score vector before we select the species to emit. We implemented three pooling rules. *Mean* averages the per-tile logits; it is robust to a single noisy tile but dilutes a species that is strong in only one or two tiles among the many background tiles where it is absent. *Max* keeps each species' single strongest tile; it suits a species confined to one tile but is fragile, since one over-confident background tile can promote a species across the whole image. *Softmax-mean* turns each tile into a normalised probability vote and averages those votes, so no single tile can dominate while evidence still accumulates across tiles. The final system uses softmax-mean. Plain mean was explored on an earlier head-only checkpoint (experiment 010) and dropped; it was not re-tested on the final model. The two-stage comparison behind this choice, with the same-checkpoint caveat, is reported in Section 4.5.

After aggregation we replace standard top-k selection with an adaptive thresholded rule: we emit every species whose probability exceeds a fixed threshold T, subject to a floor of $k_{\min}$=2 and a ceiling of $k_{\max}$=10 species per quadrat. This matches the test set's empirical cardinality of roughly two to three species per quadrat without committing to a single k. The operating point (T=0.03) and its comparison against fixed top-k, gap, and relative-threshold rules are reported in Section 4.

3.6. Calibration and Ensembling

Two further ingredients complete the inference recipe. First, to counteract the long-tail bias of the species prior we apply class-prior logit adjustment [13], $\ell'_s = \ell_s - \tau \log \tilde{\pi}_s$ with τ=0.25, where $\tilde{\pi}_s$ is the Laplace-smoothed training prior for species s; this requires larger margins for common species and smaller margins for rare ones. Second, the per-tile distributions from the 224- and 336-pixel passes are averaged in probability space before selection, a single-checkpoint resolution ensemble that recovers fine-scale evidence the native resolution drops. Both knobs are swept on the leaderboard in Section 4; here we fix the values used by the final system.

3.7. Final System

Our final system is a single BioCLIP 2.5 ViT-H/14 backbone trained under the two-stage schedule of §3.2 on the enlarged i001 manifest. Only the last four transformer blocks (plus the final layer norm and projection) are unfrozen in the second stage, and both stages share the joint species/genus/family cross-entropy loss of §3.3. At inference (Figure 5), each quadrat is partitioned into a non-overlapping 4×4 grid of sixteen tiles; every tile is forwarded through the encoder at both 224 and 336 pixels. Per-tile softmax probabilities are aggregated by species-wise mean across the sixteen tiles, the two resolutions are averaged, class-prior logit adjustment (τ=0.25) is applied, and every species whose post-adjusted probability exceeds T=0.03 is emitted, subject to $k_{\min}$=2 and $k_{\max}$=10. The two-resolution average is a single-checkpoint ensemble of the i002 model, not a multi-checkpoint ensemble; it is applied at inference time, not as part of the training-time pipeline. This configuration is our headline submission, scoring 0.41826 public and 0.40283 private Macro F1 on the PlantCLEF 2026 leaderboard; the experiments that justify each component follow in Section 4.

Beyond this anchor we submitted three architecture-orthogonal pivots that aimed to add information the visual encoder cannot read from a single tile: a triple-backbone late-fusion classifier, in which the frozen BioCLIP 2.5, DINOv3, and ConvNeXt-V2-L feature vectors are concatenated and a new species-only head is trained on the concatenation (a feature-level late fusion, not an architectural change to any backbone, and without the auxiliary taxonomy heads of the anchor); a genus and family co-occurrence rerank; and a seasonal phenology prior built from GBIF month-of-observation histograms (full method in Appendix D). None improved over the visual anchor; all three are reported in Section 4.

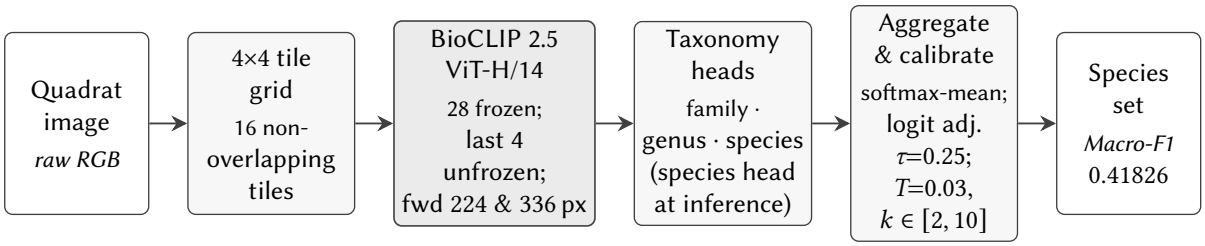

Figure 5: Inference pipeline for the ANU PlantCLEF 2026 submission (public Macro F1 0.418265). Each quadrat is partitioned into a non-overlapping 4×4 grid of sixteen 448-pixel tiles and forwarded through the partially fine-tuned BioCLIP 2.5 ViT-H/14 (lower 28 transformer blocks frozen; last 4 plus LN and projection unfrozen) at two inference resolutions, 224 and 336 pixels. The encoder produces outputs at three taxonomic levels (family, genus, species); only the species head is used at inference time. Per-tile softmax probabilities from the species head are averaged across the 16 tiles and across the two resolutions, then class-prior logit adjustment (τ=0.25) down-weights common species before an adaptive probability threshold (T=0.03, $k_{\min}$=2, $k_{\max}$=10) selects the output species set.

4. Experiments

This section justifies each component of the final system (defined top-down in §3.7) with leaderboard evidence, in dependency order: the evaluation protocol, the baseline ladder, the main supervised fine-tuning result, the unfrozen-blocks ablation, tile aggregation, calibration and ensembling, and finally the submitted configurations. Throughout, we treat this as a development trace rather than a single controlled ablation: many comparisons move more than one factor at once, and we flag the confounds where they occur. The complete per-direction trace, with exploration-depth labels, is given in Appendix A.

4.1. Evaluation Protocol

The PlantCLEF 2026 task is scored by the macro-averaged per-quadrat F1 of §1 on a hidden test set of 2,105 vegetation quadrats. Each per-quadrat F1 compares the predicted species set against the ground-truth species set for that quadrat, and those scores are then averaged within each transect and across transects. The Kaggle leaderboard reports two numbers: a *public* Macro F1 computed on a fixed subset of the test quadrats and visible during the competition, and a *private* Macro F1 computed on the complementary subset and revealed only at the close. We report both wherever a configuration was submitted, because the two do not always rank configurations the same way (§4.6); all leaderboard numbers in this paper are taken from the Kaggle API as archived in Appendix A. From repeated near-identical submissions we estimate a run-to-run leaderboard noise floor of roughly 0.002 Macro F1, and we treat differences below this as not significant.

We also track held-out top-1/top-5 accuracy on a 10% single-plant validation split for model selection. This validation signal correlates only weakly with quadrat Macro F1 ($r \approx 0.4$ across our strongest runs); we use it to choose checkpoints but not to rank final submissions, and we defer the analysis of this decoupling to §5.1.

4.2. Baselines

Before fine-tuning we measured how far an off-the-shelf BioCLIP feature space goes on this task. Zero-shot tile classification against species text prompts reached only ~0.12 public Macro F1; a frozen-backbone linear probe roughly doubled this to 0.242; a non-parametric few-shot support bank reached 0.307; and a single-head partial fine-tune (experiment 006) reached 0.222. These baselines establish that BioCLIP carries useful biological signal but that neither zero-shot matching nor a frozen backbone bridges the single-plant to multi-species shift. Full per-baseline scores and the remaining early directions are tabulated in Appendix A; we carry only the fine-tuning line forward here.

4.3. Supervised Fine-tuning Results

The decisive gains came from partially fine-tuning BioCLIP 2.5 ViT-H/14 with the taxonomy-aware multi-task head of §3.3. The first strong model, experiment 010 (shared-trunk head, five taxonomic levels, original 1.41M-image manifest, five epochs), reached a best single-configuration public Macro F1 of 0.391 (private 0.383); the four-block checkpoint we use as the reference point for the ablations below scores 0.38333. Our final model, experiment i002 (per-head MLPs, three taxonomic levels, the larger 2.65M-image i001 manifest, ten epochs), reached 0.41165 public / 0.38132 private at single-resolution inference, our best single-configuration result before calibration and ensembling.

We stress that the 010 →i002 comparison is *not* a controlled ablation: between the two we simultaneously changed the training data, the schedule (epochs and effective batch size), the head topology (shared trunk versus per-head MLPs), and the head count (five versus three). No run isolates any one of these factors. We therefore report i002 as the stronger configuration overall and make no causal claim that the per-head head design, or any single other change, produced the gain; the descriptive head-representation analysis is reported separately (§5.2).

Table 2 consolidates the BioCLIP 2.5 fine-tuning variations and the configuration axes that distinguish them, making explicit that consecutive rows differ along several axes at once.

Exp.	Manifest	Ep.	Head (levels)	Unfroz.	Pub.	Priv.
006	—	—	single linear (1)	partial	0.222	0.272
010	1.41M	5	shared MLP (5)	4–8 (swept)	0.391	0.389
i002	2.65M	10	per-head MLP (3)	4	**0.412**	**0.406**
i003	2.65M+GBIF[†]	10	per-head MLP (3)	4	0.400	0.402

Table 2
BioCLIP 2.5 ViT-H/14 supervised fine-tuning variations. "Pub." and "Priv." are the best public and best private Kaggle Macro F1 for each experiment group, on the same basis as Appendix Table 10; the two columns within a group may be drawn from different inference settings, so the single-configuration pre-calibration scores for 010 (0.391 public / 0.383 private) and i002 (0.41165 public / 0.38132 private) are the ones quoted in the running text. "Head (levels)" is the auxiliary-head topology and the number of taxonomic levels it predicts; "Unfroz." is the number of unfrozen transformer blocks (010 was swept over 0–12, peaking at eight public / four private, Table 3). The runs differ along several axes simultaneously (manifest, schedule, head topology, head count), so the table tracks the configurations rather than isolating any single factor. Experiment 006's full training manifest and schedule are not separately recorded. [†] i003 inherits the i002 schedule, head and unfrozen depth, changing only the manifest: it adds 148k GBIF images for species with fewer than 100 images and caps each species at 500.

4.4. Backbone Unfreezing

The one head-adjacent factor we swept within a single experiment is the number of unfrozen transformer blocks. Within experiment 010 (the 1.41M-image single-plant manifest, fixed optimiser and head configuration) we progressively unfroze the last 0, 4, 8 and 12 blocks (each stage warm-starting from the previous and training for five epochs), and finally attempted a full fine-tune. This is a staged warm-start sweep rather than a strictly controlled single-variable ablation, but it is the cleanest within-recipe evidence we have. Table 3 reports public and private Macro F1 under the common `grid_4x4`/softmax-mean pipeline.

Public Macro F1 rises from the head-only baseline (0.362) to a peak at *eight* unfrozen blocks (0.391) and then declines at twelve. The private split is flatter: the four-, eight- and twelve-block runs are statistically indistinguishable (0.383–0.389), so the takeaway is a moderate-unfreeze optimum rather than a single sharp peak. Both observations hold even though held-out validation top-5 keeps rising monotonically with depth. The full fine-tune performs substantially worse on both splits (0.301 public, 0.296 private): unfreezing the entire backbone trades the pretrained representation for single-plant fit faster than the leaderboard rewards. The lesson is that partial unfreezing helps but is easily overdone, and that the single-plant validation metric over-rewards capacity relative to quadrat Macro F1 (§5.1).

Unfrozen blocks	Best public F1	Best private F1	Val top-5
0 (head only)	0.362	0.376	0.921
4 blocks	0.383	**0.389**	0.932
8 blocks	**0.391**	0.383	0.936
12 blocks	0.375	0.384	0.940
32 (full fine-tune)	0.301	0.296	0.934

Table 3
Unfrozen-blocks sweep within experiment 010 (BioCLIP 2.5 ViT-H/14, 1.41M-image single-plant manifest; each stage warm-starts from the previous, block stages train five epochs). Public Macro F1 peaks at eight unfrozen blocks; the private split is flatter, with the four-, eight- and twelve-block runs within 0.006 of one another (0.383–0.389), so the optimum is a moderate-unfreeze plateau rather than a single sharp peak. Held-out validation top-5 rises monotonically through twelve blocks: the validation/leaderboard decoupling of §5.1. The full fine-tune was submitted but scored far below the plateau on both splits (public 0.301, private 0.296); its validation top-5 (0.934) had already dropped below the 12-block run and its validation loss rose after epoch 7 (overfitting), so we did not carry it forward. Each cell is the best-public or best-private selection for that stage; selection rules differ across cells.

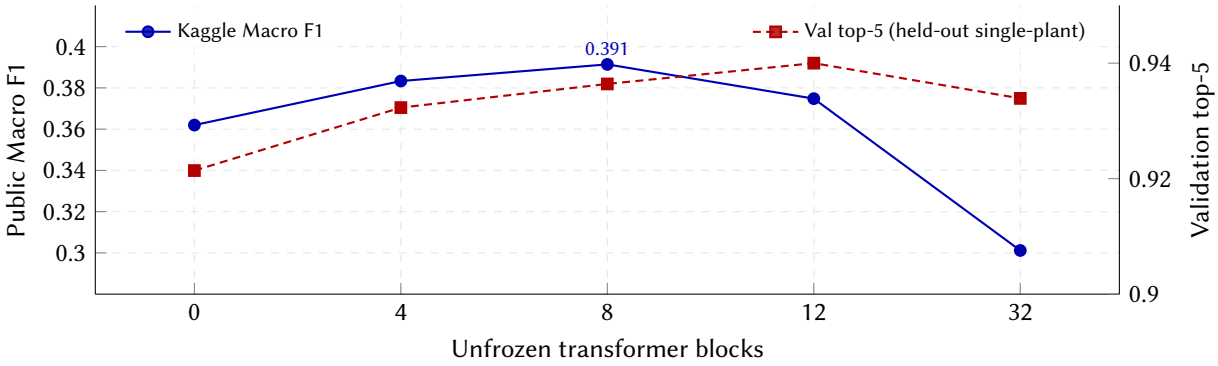

Figure 6: Unfrozen-blocks sweep visualised with both metrics. Held-out validation top-5 (dashed, right axis) rises monotonically with unfrozen depth (n=0 → 12), but leaderboard Macro F1 (solid, left axis) peaks at n=8 and declines by n=12; the full fine-tune (n=32) drops further still. The disagreement between the two curves is examined in §5.1: validation accuracy on the single-plant distribution does not predict multi-species quadrat Macro F1.

Our final model (experiment i002) repeated this staged unfreezing on its larger 2.65M-image manifest, but only as far as eight blocks (Table 4). Here the optimum is shallower: public Macro F1 peaks at *four* blocks and eight blocks is worse, so we chose the four-block checkpoint as the anchor. Because the manifest, schedule and head topology all differ from the 010 sweep, the two are not directly comparable; we read the shift only as suggestive evidence that a larger manifest reduces the useful unfreeze depth. The headline anchor uses this four-block checkpoint served as a 224+336 ensemble with logit adjustment (§4.6); this is the same *number* of unfrozen blocks as the 010 four-block row, not the identical run.

4.5. Tile Aggregation

The three pooling rules that turn the sixteen per-tile predictions into one image-level score vector are defined in §3.5 and illustrated in Figure 7: *mean*, *max*, and *softmax-mean*. Here we report which rule wins on the leaderboard, in the two stages shown in Figure 8.

The first stage is the earlier experiment 010 head-only checkpoint, where we swept all three rules at a matched tile geometry (`grid_4x4`, overlap 0.5). Because the checkpoint and grid are held fixed, this is a clean comparison, and plain mean was clearly the weakest: its best public Macro F1 was 0.25784 (private 0.32028) against 0.33830 (private 0.34039) for max and 0.34042 (private 0.37642) for softmax-

i002 stage (2.65M manifest)	Best public F1	Best private F1
head only (0 blocks)	0.394	0.388
4 blocks (anchor)	**0.412**	0.389
8 blocks	0.405	0.375
inference enhancements on the 4-block checkpoint		
+ logit adj. (τ=0.25, T=0.03)	0.406	**0.406**
+ 224+336 ensemble (anchor)	**0.418**	0.403

Table 4
i002 staged unfreeze sweep on the larger 2.65M-image manifest (ten epochs per stage, three taxonomy heads). Unlike the 010 sweep, public Macro F1 peaks at *four* blocks and eight blocks is worse, so the four-block checkpoint was submitted; the larger manifest already makes the head-only baseline strong (0.394). This sweep is partial (no twelve-block or full-fine-tune run) and is *not* directly comparable to the 010 sweep, which differs in manifest, schedule and head topology. The bottom rows show the logit-adjustment and resolution-ensemble inference steps that produce the headline anchor (§4.6).

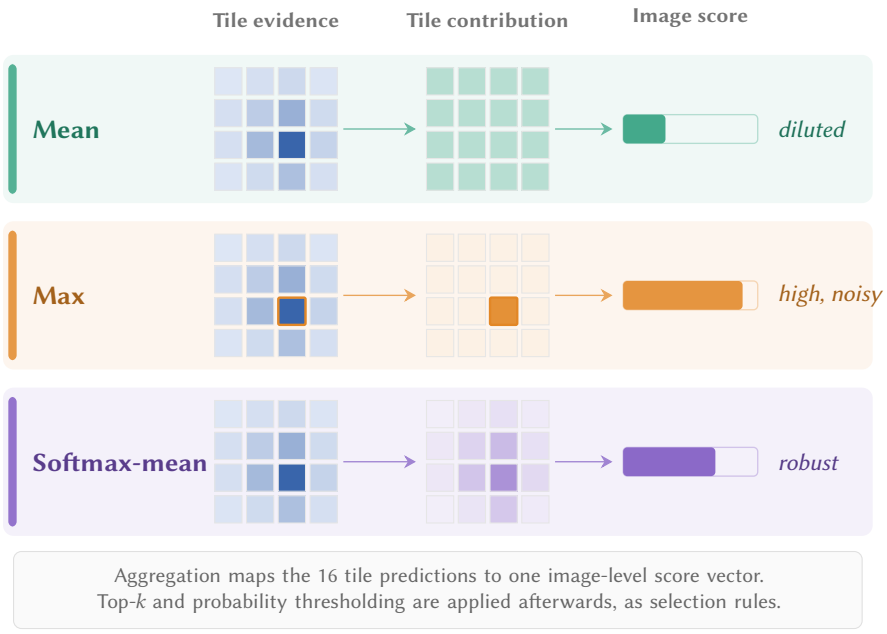

Figure 7: Inference-time tile-aggregation rules. Aggregation maps the sixteen tile-level species predictions into one image-level score vector. Mean uses all tiles equally and dilutes evidence that is strong in only a few tiles; max keeps each species' single strongest tile and is sensitive to one over-confident tile; softmax-mean averages normalised per-tile probability votes, so evidence accumulates without any one tile dominating. Top-k and probability-threshold selection are applied after aggregation and are not themselves aggregation rules.

mean, roughly 0.08 public below both alternatives. This matches the dilution argument, so we did not carry plain mean forward. We treat this as the reason mean was dropped rather than as a result on the final model, since it comes from an earlier, weaker head-only checkpoint and is confounded by checkpoint and training-data differences.

The second stage is the i002 four-block checkpoint (`grid_4x4`, no overlap, 224 pixels), where the submitted comparison was max against softmax-mean only; plain mean was not re-submitted here, so the 010 mean numbers and the i002 softmax-mean numbers are *not* a same-checkpoint comparison. Softmax-mean was again the better of the two: at each method's best selection it reached 0.41165 public Macro F1 against 0.37509 for max (+0.037), with the same ordering on the private split (0.38883 against 0.35724). We adopt softmax-mean with an adaptive probability threshold for the final system.

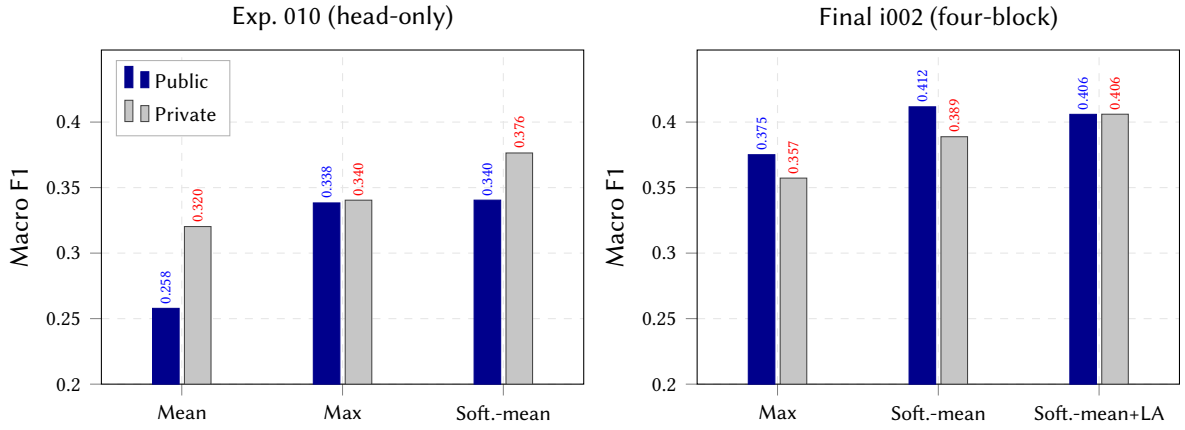

Figure 8: Two-stage tile-aggregation comparison (public and private Macro F1). The left panel is the earlier experiment 010 head-only checkpoint at matched tile geometry (`grid_4x4`, overlap 0.5), where plain mean was explored and found weakest, so it was dropped. The right panel is the final i002 four-block checkpoint (`grid_4x4`, no overlap, 224 pixels), where the submitted comparison was max against softmax-mean only; plain mean is intentionally absent from this panel because it was not re-submitted on this checkpoint. Softmax-mean improves on max on both splits, and softmax-mean with logit adjustment (τ=0.25) is the most robust on the private split. Each bar is the best-public or best-private selection for that mode.

4.6. Calibration and Ensembling

On the i002 four-block checkpoint we fixed the remaining inference knobs through leaderboard sweeps: the logit-adjustment temperature τ, the probability threshold T, the selection rule, and the resolution ensemble. Unless noted, numbers are public Macro F1 on the i002 last-four-blocks single-resolution (224-pixel) checkpoint.

Logit adjustment τ and probability threshold T. Table 5 reports the joint sweep over $\tau \in \{0.25, 0.5\}$ and $T \in \{0.03, 0.05, 0.07\}$. The optimum is sharp: τ=0.25, T=0.03 dominates every other cell by between 0.022 and 0.083 Macro F1, and Macro F1 falls off monotonically along both axes. Doubling τ from 0.25 to 0.50 at fixed T=0.03 costs 0.022, evidence that the long-tail boost saturates quickly and that pushing it harder starts emitting too many rare-but-implausible species. We stress that τ=0.25 and T=0.03 were selected by following our public-leaderboard score progression across submissions rather than on a held-out multi-species validation set, which we lacked. The coarse two-by-three grid above is therefore a leaderboard-guided operating point chosen as a reasonable balance under a limited submission budget, not the output of a rigorous hyperparameter search; a finer joint sweep with proper held-out tuning is left to future work (§6).

τ	T=0.03	T=0.05	T=0.07
0.25	**0.40590**	0.36423	0.35069
0.50	0.38440	0.34047	0.32248

Table 5
Joint $\tau \times T$ sweep on the i002 last-four-blocks checkpoint, single-resolution 224-pixel inference. Public Macro F1. The peak at (τ=0.25, T=0.03) is the operating point taken into the 224+336 ensemble that produces the headline 0.41826 in Table 7.

Selection rule. Holding τ=0.25 fixed, we compared an adaptive probability threshold (emit every s with $p_s > T$) against fixed top-k, a gap rule, and a relative threshold. Probability-thresholding at T=0.03 wins decisively (Table 6); fixed top-k is second-best at k=2, consistent with the test set's empirical mean of ~2–3 species per quadrat. The gap and relative-threshold rules underperform the anchor by between 0.06 and 0.10 Macro F1; their full sweeps are reported in Appendix F.

Resolution ensemble. Averaging the same checkpoint at two inference resolutions lifts the re-

Rule	Value	Public Macro F1
probability threshold (*anchor*)	T=0.03	**0.40590**
probability threshold	T=0.05	0.36423
probability threshold	T=0.07	0.35069
fixed top-k	k=1	0.28633
fixed top-k	k=2	0.39878
fixed top-k	k=3	0.39731
fixed top-k	k=4	0.38042
fixed top-k	k=5	0.35959

Table 6
Selection-rule ablation at τ=0.25 on the i002 last-four-blocks single-resolution checkpoint (probability-threshold and top-k families). Probability-thresholding at T=0.03 dominates; fixed top-k is competitive at k=2 because the test-set mean cardinality is empirically ∼2–3 species per quadrat. The gap and relative-threshold families (Appendix F) underperform the anchor by between 0.06 and 0.10 Macro F1.

sult substantially. Going from the best single-resolution 224-pixel softmax-mean result before logit adjustment to the 224+336 probability-averaged ensemble at τ=0.25, T=0.03 raises public Macro F1 from 0.41165 to 0.41826 (Δ=+0.0066) and private Macro F1 from 0.38132 to 0.40283 (Δ=+0.022). This delta bundles the addition of logit adjustment with the resolution ensemble; the cleaner matched-LA comparison at τ=0.25, T=0.03 (single-resolution 0.40590 from Table 5 versus ensemble 0.41826) gives a public lift of about +0.012. The private gain remains more than 3× the public gain, suggesting that the 336-pixel pass recovers rare-species pixels the single-resolution recipe drops, and those species are the ones the private split weights most heavily. This is a single-checkpoint ensemble of the i002 model, not a multi-checkpoint ensemble.

Public/private caveat. The best public and best private operating points are not the same submission. Adding logit adjustment to softmax-mean scores 0.40590 public but 0.40600 private (the strongest private result of the project), so the configuration we report as the public headline is not the one that maximises private Macro F1. We carry both columns through the final-submission table below rather than collapsing to a single number.

4.7. Final Submission

Our headline configuration is the i002 four-block checkpoint under the full inference recipe of §3.7: 4×4 grid tiling, a 224+336 resolution ensemble, softmax-mean aggregation, logit adjustment (τ=0.25), and adaptive probability-threshold selection (T=0.03, $k_{\min}$=2, $k_{\max}$=10). It scores **0.41826** public and 0.40283 private Macro F1, the best public result of the project.

Beyond this anchor we submitted three architecture-orthogonal pivots that each rerank the anchor posterior with an additional signal; Table 7 collects all four submitted configurations.

Config	Macro-F1	Δ
BC2.5 last 4 blocks, 224+336+LA, T=0.03	**0.41826**	—
+ cRT triple, w=0.3	0.38195	−0.036
+ Genus α_g=0.1	0.41246	−0.006
+ Pheno., β=1.0	0.41346	−0.005

Table 7
All four submitted configurations on the PlantCLEF 2026 hidden test set (public Macro F1). The cRT pivot is strongly negative; the genus rerank and the phenology prior move the anchor by −0.006 and −0.005 respectively, magnitudes small (about two to three times the run-to-run leaderboard noise floor we characterise at ∼0.002) but consistent in sign. We read this as the BioCLIP 2.5 posterior implicitly absorbing both signals through training-distribution co-occurrence; the orthogonality argument is developed in §5.3.

Pivot 1: triple-backbone late fusion (cRT). This is a feature-level late fusion, not an architectural

pivot: we retrained two MLP heads on a frozen 3328-d concatenation of BioCLIP 2.5 (1024-d), DINOv3-Base (768-d), and ConvNeXt-V2-L (1536-d) features (best validation accuracy 63.95%) and mixed them with the anchor at w=0.3. Both heads are flat species-only classifiers (3328→7806) and, unlike the anchor, carry *none* of its taxonomic regularization: they are supervised on species labels alone, with no auxiliary genus or family heads and no weighted joint taxonomic loss (§3.3). The two differ only in their long-tail treatment—one trained with plain cross-entropy, the other with a class-balanced sampler (the cRT recipe)—and we averaged them before the mix. A cheap pre-submission diagnostic gate (top-one agreement and submission Jaccard against the anchor) already flagged the head as structurally divergent (agreement 0.094, Jaccard 0.078). On submission it scored **0.38195** (Δ=−0.036): the divergence was diversity going *against* signal, a peakier posterior whose high-confidence wrong picks dominate the mix.

Pivot 2: genus/family co-occurrence reranking. A "siblings stick together" log-prior built from the anchor's own aggregated genus/family supports, applied at matched volume (α_g=0.1, Jaccard 0.887 with the anchor), swapped 11% of selections but moved the score by only **0.41246** (Δ=−0.006, about three times the noise floor). A prior derived from the model's own logits carries essentially no new information.

Pivot 3: seasonal phenology prior. A date-conditioned prior over GBIF day-of-year histograms, combined multiplicatively with the visual posterior at β=1.0 (full method in Appendix D), flipped 12.8% of top-one picks but scored **0.41346** (Δ=−0.005, about twice the noise floor). Date is information-theoretically orthogonal to the image, yet the gain is negligible: the visual model trained on dated images already encodes seasonal phenotype, and the GBIF effort itself is summer-biased. We retain phenology as a novelty contribution and discuss why it (and the other pivots) fail to add signal in §5.3.

5. Analysis

The experiments fix the system; this section explains why it looks the way it does, and is careful to separate what we can measure from what we can only describe. Section 5.1 documents the central decoupling of the project: held-out accuracy on the single-plant validation split barely predicts multi-species quadrat Macro F1, which reframes the bottleneck as a distribution-shift problem rather than a model-capacity one. Section 5.2 probes the trained head representations with linear CKA and taxonomic-separation diagnostics; this analysis is *descriptive*, not causal. Section 5.3 shows that variants within our own backbone have saturated into essentially one model, which both motivated the search for orthogonal evidence sources and explains its failure. Section 5.4 examines why flattening the long tail hurt rather than helped.

5.1. Validation vs Leaderboard Mismatch

A striking decoupling runs through every model-selection decision in this project: held-out accuracy on the single-plant distribution barely predicts Macro F1 on the multi-species quadrat leaderboard. Across our twelve strongest runs the two metrics correlate at only $r \approx 0.4$ (Figure 9), and within the unfrozen-blocks ablation of Table 3 they invert outright: validation top-5 accuracy rises monotonically with unfrozen capacity while leaderboard Macro F1 peaks in the middle and collapses at full fine-tune. We read this as direct evidence that the bottleneck on this benchmark is the single-plant to multi-species distribution shift, not model capacity on the single-plant distribution itself. It is also why we use validation accuracy only to choose checkpoints, never to rank final submissions (§4.1): the headline anchor does not sit at the validation maximum, and the configurations that do (the i003 capped checkpoints, the 015 PlantCLEF24+iNaturalist mix) score lower on the leaderboard.

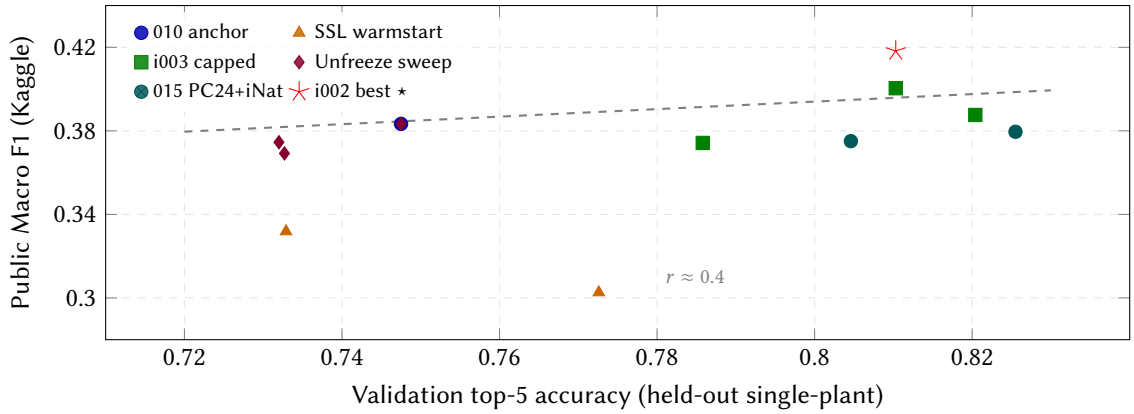

Figure 9: Validation top-5 accuracy on the held-out single-plant split versus public Macro F1 on the multi-species quadrat leaderboard, across our twelve strongest runs. The headline anchor (⋆, i002 last-four-blocks 224+336+LA ensemble) does not sit at the maximum validation point; the i003 capped checkpoints (green squares) and the 015 PC24+iNat mix (teal) reach higher validation accuracy but lower leaderboard score. The dashed line indicates the weak monotone trend ($r \approx 0.4$).

5.2. Head Design and Representation Geometry

The two supervised fine tunes wire their classification heads differently (Figure 4): experiment 010 feeds a single shared multi layer perceptron into five linear heads, whereas experiment i002 gives each of its three heads an independent perceptron. To understand what each design does to the feature space we ran a descriptive probe of the trained representations. This analysis is not causal: the two experiments also differ in training data, schedule, number of unfrozen blocks, and head count, and they use different validation splits, so we compare each representation only against its own encoder and never compute cross experiment similarity (only 62 of the 4,000 sampled images overlap between the two splits).

For a family stratified sample of 4,000 validation images per experiment we captured the BioCLIP encoder feature and every MLP hidden vector with forward hooks, then measured linear centred kernel alignment (CKA) and lightweight taxonomic separation diagnostics (cosine silhouette at the genus and family levels). The results are summarised in Table 8 and Figure 10.

Two observations are robust within each experiment. First, every MLP output stays highly aligned with its own encoder feature (linear CKA between 0.75 and 0.84), so neither head design discards the BioCLIP geometry. In i002 the three per head MLPs are distinct rather than redundant copies (pairwise CKA 0.73 to 0.84): the family MLP departs most from both the encoder (0.75) and the species MLP (0.73), while the species MLP stays closest to the encoder (0.82). Second, in both designs the trained MLP is modestly more taxonomically structured than that experiment's own raw encoder feature (for example the 010 shared trunk raises genus cosine silhouette from 0.18 to 0.21). Within i002 the genus and family MLPs separate taxa more sharply than the species MLP (family level silhouette 0.26 for the family MLP versus 0.16 for the species MLP), which is the expected signature of each MLP specialising toward its own head's loss.

We stress that these are descriptive statements about representation geometry. They characterise what the shared and per head designs do to the feature space, but they do not isolate the effect of the head topology on the leaderboard score, which remains confounded with the data and schedule differences between the two experiments. We therefore make no causal claim, and we do not treat the 010 versus i002 comparison as a controlled ablation.

5.3. Within-backbone Saturation and Orthogonality

Before pivoting to architecture-orthogonal evidence sources, we measured whether further within-backbone diversity could break the i002 anchor by comparing the four-blocks i002 checkpoint against

Exp.	Representation	CKA→enc.	Fam. silh.	Gen. silh.
010	encoder	—	0.134	0.183
010	shared MLP	0.845	0.177	0.215
i002	encoder	—	0.134	0.185
i002	species MLP	0.821	0.158	0.220
i002	genus MLP	0.806	0.193	0.252
i002	family MLP	0.749	0.260	0.296

Table 8
Within experiment representation analysis. CKA is linear centred kernel alignment of each MLP output to its own encoder feature; silhouette is cosine, at the family and genus levels, on the same 4,000-image sample. Cross experiment CKA is omitted because the two experiments use different validation splits (only 62/4,000 sampled images overlap). i002 cross head CKA: species/genus 0.84, species/family 0.73, genus/family 0.84.

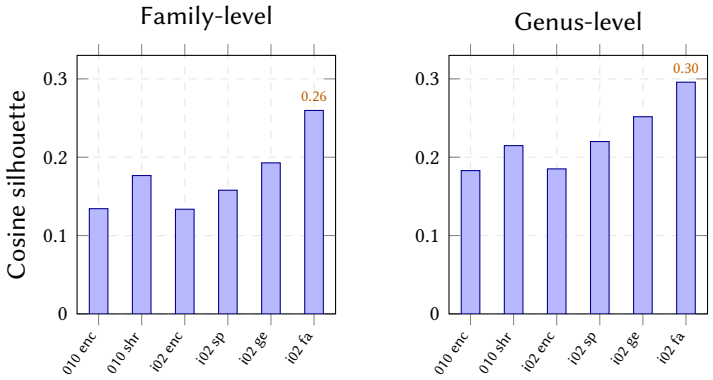

Figure 10: Taxonomic structure of each head representation on the family stratified 4,000-image samples: cosine silhouette at the family and genus levels. Within each experiment the trained MLP increases structure over its own raw encoder feature; within i002 the family and genus MLPs are the most structured (family level silhouette 0.26 for the family MLP versus 0.16 for the species MLP). Magnitudes are comparable only within an experiment, not across, because the two experiments use different validation samples.

the eight-blocks i002 checkpoint and against epoch-5 versions of both. Table 9 reports the pairwise top-one agreement and submission Jaccard on 2,105 test quadrats.

Pair	Top-1	Jaccard
blk_4 vs. blk_8	0.901	—
blk_4 vs. blk_4 (epoch 5)	0.807	0.664
4-way intra-BC2.5 ens. vs. anchor	—	0.923

Table 9
Within-backbone diagnostic gates. Top-1 agreement on a 7,806-way classifier and submission Jaccard at the operating threshold.

The 0.923 Jaccard between the four-way intra-BioCLIP 2.5 ensemble and our anchor submission shows that all within-backbone variants are essentially one model on the test set: shuffling the unfrozen depth or the training-epoch index does not move the prediction distribution. This is what motivated the three architecture-orthogonal pivots (§4.7): if the remaining headroom is not inside the backbone, it must come from a signal the backbone cannot read.

The pivots' near-zero deltas (Table 7) reveal an under-appreciated property of strong fine-tuned visual classifiers on biological data: *information-theoretic orthogonality is not the same as statistical orthogonality* once the visual model has been trained on data where the auxiliary signal is naturally correlated with image content. cRT was structurally orthogonal but introduced wrong evidence and

dragged the anchor down by 0.036; genus reweighting was statistically dependent on the BioCLIP 2.5 posterior itself and moved the score by −0.006, close to the noise floor; the seasonal phenology prior, the most genuinely orthogonal source we tested (since date is not in the image at all), moved the score by only −0.005, again close to the noise floor, because the seasonal structure was already implicit in the leaf and flower phenotypes seen during training. The practical reading is that real headroom on this benchmark requires structurally stronger *visual* models rather than post hoc reweighting with auxiliary priors (§6).

5.4. Long-tail Behaviour

The cleanest near-controlled counterfactual in the project concerns the long tail. Experiment i003 takes the i002 manifest, augments the tail with 148,063 GBIF-sourced images (of 161,686 candidates, after dropping 13,623 that failed image decoding) for the 2,020 species with fewer than one hundred training examples, and then caps every species at five hundred images, flattening the head of the distribution (Figure 11). Because i003 changes only the training manifest relative to i002, the leaderboard outcome is unusually interpretable, and it is a surprise: this flatter distribution *hurt*, with the i003 best reaching 0.40041 public Macro F1 against 0.41826 for the corresponding uncapped i002 checkpoint. We read this as direct evidence that, on this benchmark, the marginal value of additional head-class examples is positive provided their label quality is high: capping discards information rather than rebalancing it. The augmentation did not lift any species above the hundred-image floor, so the long tail itself remained long (Figure 11, left bins); the cap only trimmed the head. Full manifest statistics are given in Appendix C.

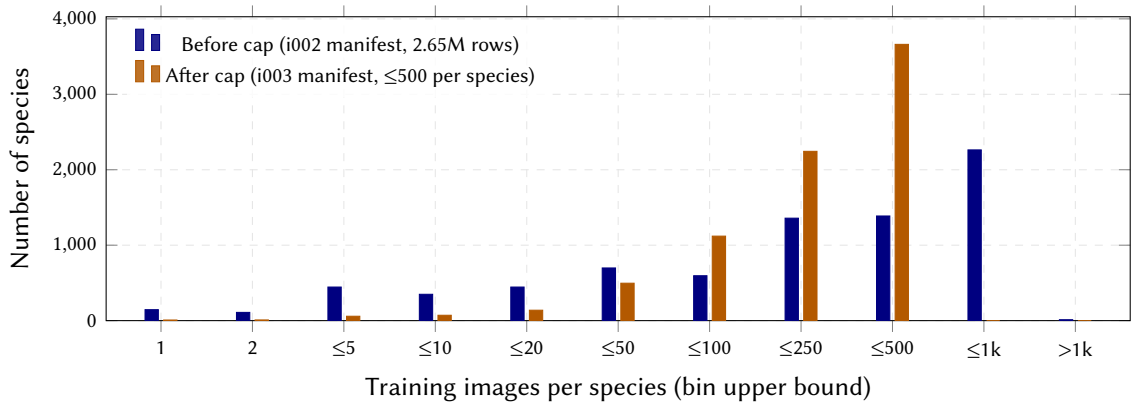

Figure 11: Species count per training-image bin, before and after the i003 per-species cap at 500. The cap collapses the two heaviest bins (≤1k and >1k) into the ≤500 bin and lifts middle bins, but does not move species out of the long tail (left bins). This flatter distribution *hurt* the public Macro F1.

6. Discussion and Limitations

Our analysis (§5) leaves a clear practical implication: real headroom on this benchmark requires structurally stronger *visual* models (end-to-end fine-tuning that preserves the prior, multi-crop test-time augmentation, test-time training, synthetic mosaic training distributions) rather than post hoc reweighting with auxiliary priors. The remainder of this section is candid about what our results do and do not establish.

What we did not cleanly ablate. Our strongest model, i002, differs from the earlier 010 fine-tune along several axes at once: the training data (1.41M → 2.65M images, a different manifest), the schedule (five → ten epochs and a larger effective batch), the head topology (a shared MLP trunk → independent per-head MLPs), and the head count (five → three taxonomic levels). No run isolates any one of these

factors, so we report i002 as the stronger configuration overall and make *no* causal claim that the per-head head design caused the gain. Nor do we ablate the taxonomy supervision itself: we never trained a species-only control with the auxiliary heads removed, so the auxiliary losses are reported as a deliberate regulariser, not as a measured improvement. The one head-adjacent factor we did isolate, the number of unfrozen blocks (holding the 010 head fixed), shows a sharp optimum at four blocks (§4.4), and that is the surface the final anchor uses. The single near-controlled counterfactual we have is the i003 cap (§5.4), where only the manifest changed and the score dropped.

Is the four-block recipe specific to BioCLIP? Whether the four-block optimum of §4.4 is BioCLIP-specific we cannot settle empirically, since we never fine-tuned an alternative backbone at varying depth (DINOv3 and ConvNeXt-V2 entered only frozen, in the cRT pivot); but the result separates into two parts. That a full fine-tune on a long-tailed, distribution-shifted target underperforms a partial unfreeze is a generic property of partial fine-tuning (§2, §5.1) and should recur on any strong backbone. The steepness of our collapse (a plateau near 0.39 down to 0.301) is more plausibly BioCLIP-specific: its prior is aligned to the Linnaean tree of life this task scores against, so overwriting it is maximally costly, whereas a self-supervised backbone like DINOv2 is less directly tied to the species labels and may be more robust to full fine-tuning. The organiser-provided DINOv2, distributed both frozen and fully fine-tuned on PlantCLEF [8], is the natural testbed; we read the qualitative pattern as likely to transfer while treating the exact optimal depth as a per-backbone hyperparameter, and leave the sweep to future work.

Public and private leaderboard instability. The public and private splits do not rank our configurations the same way. Our best public submission (0.41826) is not our best private one: the lightly logit-adjusted single-resolution run is the strongest on the private split (0.40600), and an i002+010 cross-checkpoint ensemble is near-best private while weak public. The three pivot deltas (-0.005 to -0.036) are mostly within or near the ~ 0.002 run-to-run noise floor we characterise in §4.1, so apart from cRT they should be read as "no usable effect" rather than as measured decreases. We therefore carry both columns throughout the paper and avoid implying that the best public run is best overall.

Limitations of the representation analysis. The head-representation study (§5.2) is descriptive geometry, not a controlled experiment. It uses a single checkpoint per design, a single 4,000-image sample per experiment, and two different validation splits (only 62 of 4,000 sampled images overlap), so we compare each representation only against its own encoder and never cross-experiment. It characterises what the shared and per-head designs do to the feature space; it does not isolate the effect of head topology on Macro F1. More broadly, the full development trace in Appendix A is exactly that, a trace rather than a controlled ablation table, and most of its rows change several factors at once.

Future work. Several directions follow directly from these findings. The first is to pull on the data axis rather than the capacity axis: a synthetic mosaic training regime that composes single-plant images into multi-species scenes with controlled overlap would let the network see the quadrat distribution during training, which neither the PlantCLEF 2024 corpus nor iNaturalist habitat shots do at scale. The second is self-distillation against our own production checkpoint, using its predictions on the unlabelled LUCAS pseudo-quadrat corpus as soft targets filtered by confidence and entropy. The third is genuinely orthogonal evidence that is not in the image at all, in particular a location-conditioned bioclimatic envelope prior built from training-image coordinates against WorldClim variables at each test site, the one orthogonal axis that visual features cannot recover from the image alone (this experiment is blocked here only by the absence of test-site coordinates). Finally, per-class threshold calibration on a held-out multi-species validation set would target Macro F1 more directly than the global probability threshold we currently use, and could absorb residual species-frequency imbalance that our manifest leaves behind.

Two inference-time knobs we fixed pragmatically also deserve a proper study. The calibration parameters (τ=0.25, T=0.03) were selected by following our public-leaderboard score progression under a limited submission budget, not by held-out tuning; a rigorous joint hyperparameter sweep with a multi-species validation split would establish whether this operating point is genuinely optimal or merely adequate. Likewise, the 4×4 tiling at $\{224, 336\}$ pixels was an initial trade-off we never ablated on the final checkpoint: because a coarse grid can down-sample small plants while a finer grid both fragments individual specimens across tile boundaries and multiplies inference cost quadratically, a systematic sweep over grid density and tile size (ideally with overlap and multi-scale tiling, as the phenology pivot already used) is required to confirm the trade-off rather than assume it.

7. Conclusion

We presented the ANU team's submission to PlantCLEF 2026, which placed **7th** on the official leaderboard under the team name **ARM Wision**. Our headline system, specified in full in §3.7, is a partial fine-tune of BioCLIP 2.5 ViT-H/14 with only the last four transformer blocks unfrozen, trained on the enlarged PlantCLEF 2024 + iNaturalist (i001) manifest with taxonomy-aware auxiliary heads and evaluated under tiled softmax-mean inference with class-prior logit adjustment and a 224+336 resolution ensemble. It reached a public Macro F1 of **0.41826** and a private Macro F1 of **0.40283**.

Three findings carry the bulk of what this benchmark taught us. The first is that unfreeze depth matters on BioCLIP 2.5: a moderate depth of four to eight blocks is the operating region (the deployed model uses four), while a full fine-tune destroys the Tree-of-Life prior and drops the model to 0.301 public / 0.296 private. The second is the validation versus leaderboard decoupling: across our twelve strongest runs the two metrics correlate at only $r \approx 0.4$, and within the unfreeze ablation they invert entirely, which is direct evidence that the bottleneck on this benchmark is the single-plant to multi-species distribution shift rather than model capacity. The third is that auxiliary post-hoc reweighting of an already strong visual posterior carries essentially no usable signal beyond what the visual model has already absorbed: a triple-backbone late-fusion classifier (Δ=−0.036) introduced wrong evidence and hurt, while genus/family co-occurrence (Δ=−0.006) and a seasonal phenology prior (Δ=−0.005) both moved the anchor by amounts close to our run-to-run noise floor. Fine-tuned visual classifiers on biological data absorb these signals implicitly through training-distribution co-occurrence, leaving little for post-hoc combination to exploit.

We close by stressing what these results do *not* establish: the 010 →i002 gain is confounded across data, schedule, and architecture, and we make no causal claim for any single change. We then point to §6 for the limitations and the directions for future work that follow from these findings.

A. Experiment Summary and Development Trace

This appendix collects every major direction we explored for PlantCLEF 2026, what each one tested, how deeply we pursued it, its best leaderboard scores, and what we took away from it. It is intended as an honest development trace rather than a controlled ablation study. Over the project we made more than six hundred Kaggle submissions; the tables below summarise them by experiment or by inference direction rather than listing every run. No score is inferred from a training log. Where a private score could not be recovered we write `[nf]` (not found), and directions that never produced a submission (data construction and diagnostic analyses) are marked `[ns]`.

How to read "exploration depth". We label each direction with one of five depths. *Deep* means many runs or sweeps, follow-up analysis, or a component of the final system (for example the i002 model, the aggregation choice, and the headline ensemble). *Moderate* means several meaningful runs but not an exhaustive sweep (for example the early baselines). *Shallow* means one or a few quick trials. *Diagnostic only* means an analysis used to understand behaviour rather than to chase a score. *Dropped*

Table 10
Model and training experiments. Public/private are the best Kaggle Macro F1 for that experiment group; for the model rows the private value is the best private of the group. Depth: Deep / Moderate / Shallow / Diagnostic / Dropped. Scores are not a controlled ablation; see the caveats in Section A.7.

Exp.	Direction / main change	Pub.	Priv.	Depth	Fwd?	Main lesson
001/002	BioCLIP zero-shot tiling (no training)	~0.12	~0.12	Mod.	concept	Biologically useful but far too weak zero-shot; needs a trained head.
003	BioCLIP 2.5 linear probe (frozen)	0.242	0.281	Mod.	No	Doubles zero-shot but plateaus; frozen backbone cannot bridge the domain shift.
004	BioCLIP few-shot support bank	0.307	0.296	Mod.	No	Prototypes beat the probe but stay well short of fine-tuning.
005	DINOv3 SSL + fine-tune	0.140	0.142	Shal.	No	DINOv3 SSL-adapt scored far below BioCLIP fine-tunes.
006	BioCLIP 2.5 partial fine-tune (linear head)	0.222	0.272	Mod.	partial	Unfreezing raised val but not public; needed a richer head + better inference.
007	BioCLIP zero-shot + SAM vegetation mask	0.140	0.142	Drop.	No	SAM-weighted aggregation did not improve on plain zero-shot (~70 h inference).
008	DINOv3+PlantNet fine-tune (no local folder)	0.375	0.351	Mod.	No	Strongest non-BioCLIP line, still below 010/i002.
009	BioCLIP 2.5 + PlantNet data	0.208	0.248	Shal.	No	PlantNet data worse than PlantCLEF 2024 single-plant.
015	PlantCLEF24 + iNaturalist mix (no folder)	0.380	0.342	Shal.	No	Higher val, lower public: a validation/leaderboard mismatch case.
010	BioCLIP 2.5 end-to-end multitask, *shared* MLP, 5 heads	0.391	0.389	Deep	Yes	End-to-end multitask jumps to ~0.39; softmax-mean + last-4-blocks are load-bearing.
011	010 + SimSiam SSL on pseudo-quadrats	0.377	0.356	Mod.	No	SSL hurt the frozen-head stage and never beat 010.
014	Unfrozen-blocks sweep (no folder)	0.375	0.343	Diag.	informs	Confirms an interior optimum in unfrozen-block count.
i001	Data download / manifest construction	[ns]	[ns]	Deep	Yes	Larger clean genus/family-filled manifest enabled the i002 jump.
i002	**BioCLIP 2.5, *per-head* MLPs, more data, no cap**	**0.412**	**0.406**	Deep	**Yes**	Final model: more clean data + per-head MLPs + softmax-mean give the project best.
i003	i002 + 148k GBIF images for <100-image species, 500/species cap	0.400	0.402	Deep	No	Capping/flattening hurt public; more clean data beat a balanced smaller set.

means a direction we explored and then abandoned after weak or near-neutral results. These labels describe effort and role, not statistical significance.

A.1. Baselines and zero-shot experiments

We first established that the BioCLIP family of foundation models carried useful biological signal, then measured how far that signal went without task-specific training. Zero-shot tile classification (experiments 001 and 002), in which test tiles are matched against species text prompts, scored only

Table 11
Inference, aggregation, ensemble, pivot, and diagnostic directions, all built on the i002 (or 010) checkpoint. Public/private are the best Kaggle Macro F1 for the named configuration. The 224+336 ensemble is the reported headline; the best private overall is the i002 logit-adjusted run. Diagnostics produced no submission ([ns]). This is a development trace, not a controlled ablation; see Section A.7.

Item	Direction / main change	Pub.	Priv.	Depth	Fwd?	Main lesson
Agg.	softmax-mean vs max vs mean (pooling)	0.412	0.406	Deep	Yes	softmax-mean beats max (+0.037) and avoids mean dilution; the production aggregator.
mean	plain logit-mean (010 head-only only)	0.258	0.320	Drop.	No	Dilutes a species strong in 1–2 tiles across background; dropped before i002.
LA	logit adjustment τ sweep	0.406	**0.406**	Deep	Yes	Mild τ=0.25 lowers public slightly but gives the best private; $\tau \geq 0.5$ harmful.
Sel.	selection rule (probT / top-k / relT / gap)	0.406	0.406	Deep	Yes	Probability thresholding (T=0.03) at τ=0.25 beats top-k/gap/relT (no-LA peak is 0.412 at T=0.05).
Ens.	**224+336 resolution ensemble (headline)**	**0.418**	0.403	Deep	**Yes**	Multi-resolution average of one checkpoint: best public overall.
Ens.2	i002 + 010 ensemble	0.389	0.405	Shal.	No	Hurt public, near-best private: vivid public/private disagreement.
Piv.1	cRT triple-backbone late fusion	0.382	[nf]	Drop.	No	A different classifier diverged into wrong answers, not productive diversity.
Piv.2	genus/family co-occurrence rerank	0.412	0.399	Drop.	No	Priors from the model's own posterior add no information; near-neutral.
Piv.3	seasonal phenology prior (novelty)	0.413	0.388	Deep	novelty	Date is orthogonal in theory but not statistically; visual model already absorbs season.
Diag.1	head-architecture audit (010 vs i002)	[ns]	[ns]	Diag.	report	No clean ablation isolates head topology or taxonomy supervision.
Diag.2	head-feature CKA analysis	[ns]	[ns]	Diag.	report	Per-head MLPs are distinct, not redundant; family MLP most taxonomy-structured.
Diag.3	within-backbone saturation gates	[ns]	[ns]	Diag.	informs	All within-i002 variants are essentially one model; headroom needs orthogonal signal.

around 0.12 public Macro F1 at best, and the SAM vegetation-weighted variant (007) did not improve on plain zero-shot aggregation while costing roughly seventy hours of inference. A frozen-backbone linear probe (003) roughly doubled the zero-shot score to 0.242 public but then plateaued, and a non-parametric few-shot support bank (004) reached 0.307. The lesson from this group is consistent: BioCLIP embeddings are a strong starting point, but neither zero-shot matching nor a frozen backbone is enough to bridge the shift from single-plant training images to multi-species quadrats.

A.2. Supervised fine-tuning experiments

The decisive gains came from partially fine-tuning BioCLIP 2.5 ViT-H/14. A single-head partial fine-tune (006) improved validation accuracy but not public Macro F1, which signalled that we needed both a richer head and a better inference pipeline. Experiment 010 introduced the end-to-end multitask design: a shared multi-layer perceptron feeding species, genus, family, order, and class heads, with fixed auxiliary loss weights. It reached 0.391 public and made two choices that carried through the rest of the project, namely softmax-mean tile aggregation and unfreezing only the last few transformer blocks. Experiment 011 added SimSiam self-supervised pretraining on pseudo-quadrats, but this hurt the frozen-head stage and never beat 010. The strongest model, i002, kept the 010 recipe but trained on the larger and cleaner i001 manifest (about 2.65M images), switched to independent per-head multi-layer perceptrons for species, genus, and family, and trained for ten epochs. Its best single-configuration public score was 0.412 and it produced the best private score of the whole project, 0.40600, under a lightly logit-adjusted configuration. Experiment i003 then capped the data at five hundred images per species and added extra GBIF images; this flatter distribution *lowered* public Macro F1 to 0.400, so it was not carried forward. We emphasise that the gap between 010 and i002 mixes several changes (head topology, data, head count, epochs, and batch size) and is therefore not attributable to any single factor.

A.3. Inference and aggregation experiments

On a fixed checkpoint we explored how to turn sixteen per-tile predictions into one species set. Among the three pooling rules, softmax-mean (averaging per-tile softmax probabilities) was the clear choice: on the i002 four-block checkpoint it beat per-species max by about 0.037 public Macro F1, with the same ordering on the private split. Plain logit-space mean was explored only on the weaker 010 head-only checkpoint, where it was consistently the worst of the three (best public 0.258), because averaging across sixteen cells dilutes a species that is strong in only one or two tiles; we therefore dropped it before i002. For species selection, an adaptive probability threshold ($T \approx 0.03$ to 0.05) dominated fixed top-k, relative threshold, and gap rules. A mild class-prior logit adjustment (τ=0.25) traded a little public score for the best private score. Finally, averaging the same checkpoint at two inference resolutions (224 and 336 pixels) produced our headline public result of 0.418. A cross-checkpoint ensemble with 010 instead hurt public but reached near-best private (0.405).

A.4. Data experiments

Experiment i001 rebuilt the training corpus by re-downloading PlantCLEF 2024, deduplicating an iNaturalist research-grade dump, and adding GBIF images for under-represented species, then back-filling genus and family for every row. The resulting larger, cleaner manifest is what the winning i002 model trained on, with no per-species cap applied. Two data variations were tested and rejected on the leaderboard: mixing in iNaturalist data (group 015) raised validation but lowered public Macro F1, and capping at five hundred images per species (i003) also lowered public score. The extra GBIF images did not move any species above the hundred-image threshold, so the long tail remained long. The takeaway is that more clean data helped, but artificially balancing or expanding the distribution did not.

A.5. Diagnostic analyses

Three analyses were run to understand the system rather than to chase a score. A head architecture audit documented that 010 uses a shared trunk while i002 uses independent per-head multi-layer perceptrons, and established that no clean ablation isolates the head topology or the taxonomy supervision, so neither can be claimed as the cause of i002's gain. A representation analysis (linear CKA and silhouette on validation features) showed that the i002 per-head perceptrons are genuinely distinct rather than redundant, with the family head most taxonomy-structured, while all heads retain the BioCLIP geometry. A within-backbone saturation diagnostic showed that variants of i002 (different resolutions and unfreeze depths) agree on roughly eighty to ninety percent of top-1 predictions, so the remaining internal diversity is below the leaderboard noise floor; this motivated the search for orthogonal signals in the three pivots.

A.6. Final system lineage

The final system emerged along a clear path, which we describe as a sequence of choices that coincided with our best results rather than as a chain of proven causes. Zero-shot and frozen-backbone baselines established that BioCLIP was useful but insufficient. End-to-end multitask fine-tuning in experiment 010 produced the first strong model and fixed two recurring ingredients, softmax-mean aggregation and last-few-block unfreezing. The i001 data rebuild and the move to per-head perceptrons in i002 (trained on more clean data with no cap) gave the strongest checkpoint. On top of that checkpoint, the aggregation study selected softmax-mean over max and dropped plain mean, the selection study chose adaptive probability thresholding, light logit adjustment (τ=0.25) gave the best private score, and a 224+336 resolution ensemble gave the best public score of 0.41826. The three pivots (a triple-backbone late-fusion classifier, a genus/family rerank, and a seasonal phenology prior) were all reranking attempts on this anchor; none improved it, and we report the phenology prior as a novelty contribution on its own terms.

A.7. Important caveats for interpreting the appendix

- This is a development trace, not a controlled ablation table. Most rows change several factors at once (data, architecture, schedule, and inference), so a score difference between two rows cannot be attributed to any single change.
- In particular, the 010-to-i002 improvement is confounded by the data rebuild, the head redesign, the head count, the epoch count, and the batch size. We did not run a controlled shared-versus-per-head comparison, nor a species-only run that removes the taxonomy heads, so we report these as design choices, not causes.
- Public and private leaderboards disagree. Our best public configuration (0.41826) is not our best private one (0.40600), and the i002+010 ensemble is near-best on private while weak on public. We report both columns and avoid implying that the best public run is best overall.
- Held-out single-plant validation accuracy correlates only weakly with multi-species quadrat Macro F1 (about $r \approx 0.4$ across our strongest runs). The four-block checkpoint beats the eight-block one on the leaderboard despite lower validation accuracy, so validation numbers should not be over-read.
- Some directions are shallow (one or a few submissions) or have no recoverable local code, especially the manifest-only groups 008, 009, 014, and 015. Their setup is inferred from submission messages and should not be over-interpreted.
- The diagnostic analyses are descriptive. They characterise the model and its representations but do not establish causal effects on the score.
- Missing private scores are marked `[nf]` and data-only or diagnostic rows are marked `[ns]`. Several near-neutral pivot deltas (about -0.005 to -0.006) are small, about two to three times the leaderboard noise floor we estimated at roughly 0.002, and should be read as "no useful effect" rather than as a measured decrease.

B. Training Configuration of the i002 Model

The i002 model is the second stage of a two-stage schedule on the i001 manifest of §3.1. Stage 1 trains the head MLPs and the species, genus and family classifiers with the BioCLIP 2.5 backbone fully frozen; Stage 2 resumes the model weights only and additionally unfreezes the last four transformer blocks together with the final layer norm and projection. Both stages share the data, loss, optimiser family, schedule, precision, and augmentation pipeline; they differ only in the trainable surface and in the effective batch size used. Table 12 gives the full specification.

Code and artifacts. The training, inference, and submission-generation code for the system described in this paper is released at https://github.com/arm-wision/PlantCLEF2026, together with the data-manifest build scripts, the tiled-inference and calibration pipeline, and the configuration files for the i002 model. The fine-tuned backbone is the public BioCLIP 2.5 ViT-H/14 checkpoint [4] (hf-hub: imageomics/bioclip-2.5-vith14); combined with the released code and the specification in Table 12, this is sufficient to reproduce the reported submissions.

Parameter	Stage 1 (head only)	Stage 2 (last-4 blocks)
Trainable surface	head MLPs + classifiers	+ last 4 transformer blocks + `ln_post` + `proj`
Initialisation	random head, frozen backbone	Stage 1 best checkpoint (weights only)
Epochs	10	10
Per-GPU batch size	512	128
Gradient accumulation	2	4
Effective batch	2048	1024
Total optimisation steps	11,680	23,350
Warmup steps (1 epoch)	1,168	2,335
Shared across both stages		
Backbone	BioCLIP 2.5 ViT-H/14 (hf-hub:imageomics/bioclip-2.5-vith14)	
Head MLP (per task)	LayerNorm → Linear(1024→1024) → GELU → Dropout(0.2)	
Classifiers	species (7,806), genus (1,446), family (181)	
Loss	$\mathscr{L}_{\text{species}} + 0.30\,\mathscr{L}_{\text{genus}} + 0.15\,\mathscr{L}_{\text{family}}$, cross entropy with label smoothing 0.1	
Missing taxonomy labels	encoded as -1 and masked out of the auxiliary loss	
Optimiser	AdamW; head LR $1{\times}10^{-4}$, backbone LR $1{\times}10^{-6}$, weight decay $1{\times}10^{-4}$	
LR schedule	linear warmup for one epoch, then cosine decay to 1% of peak	
Gradient clip	1.0 (global norm)	
Precision	bfloat16 AMP, no `GradScaler`	
Hardware	2× NVIDIA RTX 5090, DDP via torchrun--nproc_per_node=2	
Training image size	224×224	
Train augmentations	RandomResizedCrop (scale 0.5−1.0, bicubic); HFlip p=0.5; VFlip p=0.2; rotation ±20°; ColorJitter (brightness/contrast/saturation = 0.2, hue = 0.03); RandomGrayscale p=0.05; CLIP normalisation	
Validation transform	resize shortest side to 256, CenterCrop 224×224, CLIP normalisation	
Train / val split	stratified per species, 10% to validation, seed 42, species with <5 images held in train	

Table 12
Complete training-configuration specification for the i002 model. The two rows above the horizontal divider differ between the two stages; everything below the divider is identical across stages. Numbers in the "Total optimisation steps" and "Warmup steps" rows correspond to the two-GPU run on the train split of 2,391,457 images (Table 1). The head architecture and loss weights are also discussed qualitatively in §3.3.

C. Long-tail Rebalancing: More Data vs. Per-Species Capping

The PlantCLEF 2024 single plant corpus is heavily skewed. Of the 7,806 species, the most represented twelve carry more than one thousand training images while one hundred forty five species carry a single image and four hundred forty five carry between three and five. Vanilla shuffled training thus exposes the network to roughly two orders of magnitude more head class examples than tail class examples per epoch. Experiments i002 and i003 attack this imbalance along two distinct axes, the second building directly on the first.

Experiment i002 – more data, no cap. Experiment i002 ports the partial unfreeze recipe onto the redownloaded i001 manifest, which carries 2,653,781 rows across 7,806 species (against 1,408,033 rows in the original manifest) with pre-filled genus and family labels for every row. No per species cap is applied; the rebalancing here is implicit, coming from the broader observation pipeline rather than the tail. The stratified 10% validation split yields 2,391,457 training rows and 262,324 validation rows. The downstream effect on validation accuracy is substantial: top 5 accuracy on the held out split rises from 0.9214 (010 head only) and 0.9323 (010 last four blocks) to 0.9505 (i002 head only), 0.9583 (i002 last four blocks), and 0.9615 (i002 last eight blocks).

Experiment i003 – GBIF tail augmentation with a 500-image cap. Experiment i003 builds directly on i002 by adding GBIF-sourced images for under-represented species and then capping every species at five hundred images. We queried the GBIF occurrence API for additional records of the 2,020 species that had fewer than one hundred images in the i002 manifest, yielding 161,686 candidate images; 13,623 failed a full Python Imaging Library decode (truncated, oversized, or non-image files) and were dropped, leaving 148,063 validated GBIF images that were merged into the i002 manifest. The combined manifest was then deduplicated by image path and capped at five hundred images per species through a seeded random subsample, applied once offline at manifest construction time rather than at every epoch. The cap trims the heaviest bins (the twelve species above one thousand images and the 2,263 species in the 501 to 1,000 image bin both collapse into the 251 to 500 bin) and produces a flatter distribution with 2,223,516 training rows. No new species are introduced; all 2,020 GBIF-supplemented species already exist in the i002 manifest, so the augmentation lifts only the head of the tail, not its left edge. The leaderboard outcome of this capped, GBIF-augmented manifest (a public Macro F1 of 0.40041 against 0.41826 for the uncapped i002 checkpoint) is analysed in §5.4, together with the before/after distribution (Figure 11).

D. Seasonal Phenology Prior: Full Method

The seasonal phenology prior is one of three architecture-orthogonal pivots (§3.7); on the leaderboard it scored below the visual anchor, and its empirical result is reported in Section 4. We retain the full method here as a self-contained novelty contribution.

Each test quadrat carries an observation date in its identifier (e.g. `CBN-PdlC-A1-20130807` → 7 Aug 2013); the visual backbone never sees this. Date is therefore the cleanest source of information that is *information-theoretically orthogonal* to the image-based posterior. We formalise a phenological prior as a circular probability density over day-of-year, estimated nonparametrically from GBIF month-of-observation histograms and combined with the visual posterior multiplicatively in log space.

Formulation. For a species s observed on day-of-year $d \in \{1, \dots, 365\}$, let $P(s \mid d)$ denote the species' observability density. We define the combined posterior as

$$\log p_{\text{final}}(s \mid I, d) \;=\; \log p_{\text{vis}}(s \mid I) \;+\; \beta \log P(s \mid d), \tag{2}$$

where p_{vis} is the BioCLIP 2.5 anchor posterior and β is a mixing weight: β=0 recovers the unmodified visual posterior; large β collapses the prediction toward the seasonal prior alone.

Pdf estimation. We estimate the per-species pdf nonparametrically from GBIF month-of-observation histograms. The histograms are obtained via a single faceted query per species which returns a 12-bin month histogram in one request. We then build a 365-day pdf via circular Gaussian smoothing with $\sigma = 18$ d centred on each month's mid-point, and mix in a small uniform mass to prevent hard zeros in tail months:

$$P(s \mid d) \;=\; (1-\epsilon)\,\frac{1}{Z_s}\sum_{m=1}^{12} c_{s,m}\,\mathcal{N}_{\text{circ}}(d \mid \mu_m, \sigma) \;+\; \frac{\epsilon}{365}, \tag{3}$$

with ϵ=0.05 and $c_{s,m}$ the GBIF count for species s in month m.

Coverage. We fetched histograms for all 7,796 species with valid GBIF identifiers (10 of 7,806 lack a GBIF mapping). Two passes were required: a first pass at 12 worker threads (3 req/s, 21% rate-limit drops), then a retry pass at 4 worker threads (GBIF aggressively throttled a saturated client). Final coverage: **6,957 / 7,806 species (89.1%)**; the remaining 849 species fall back to a uniform DOY pdf and contribute no phenological signal.

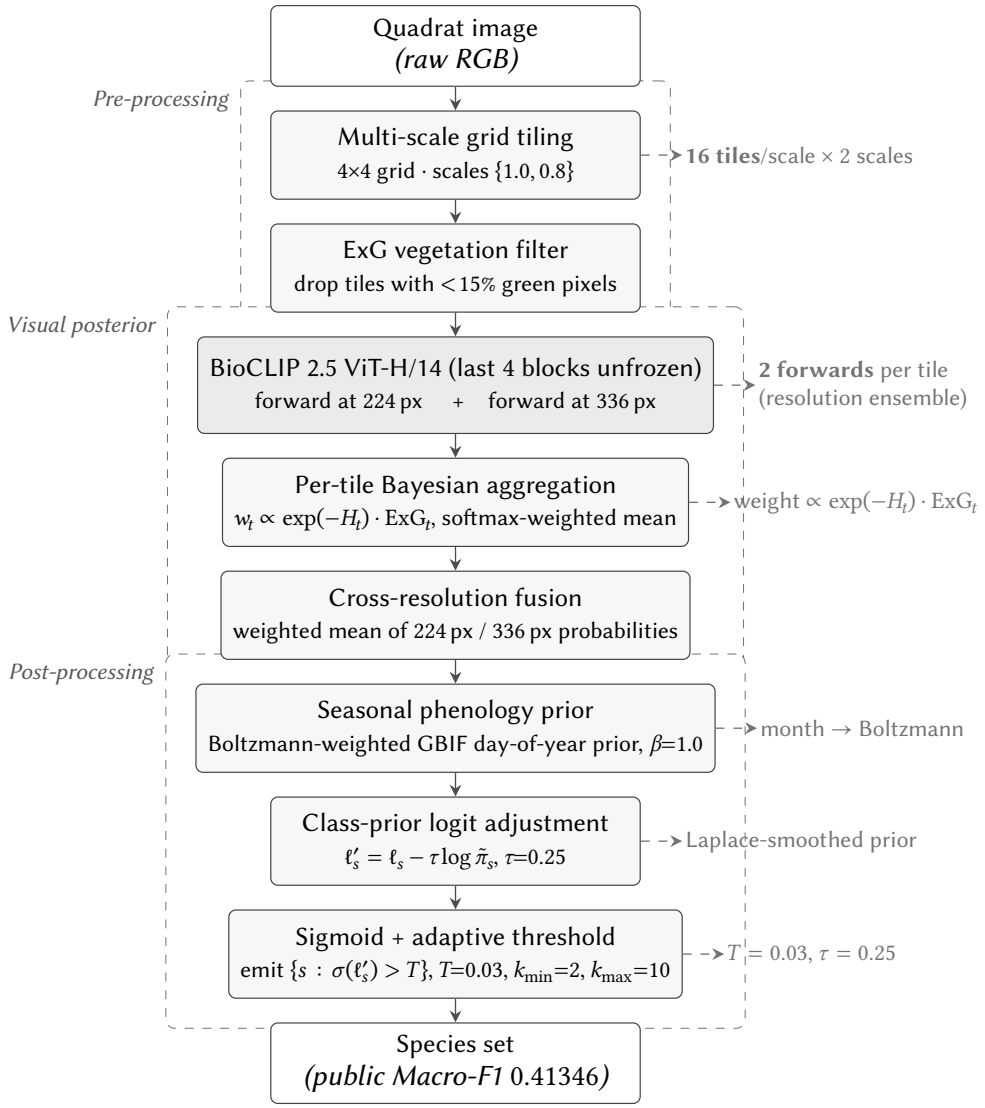

Figure 12: Inference pipeline for the seasonal phenology submission (Pivot 3, public Macro F1 0.41346). Each quadrat is processed at two grid scales (1.0 and 0.8); tiles are screened by an Excess Green (ExG) vegetation filter; the surviving tiles are forwarded through the partially fine-tuned BioCLIP 2.5 ViT-H/14 at two inference resolutions (224 and 336 pixels). Per-tile probabilities are pooled by entropy-weighted Bayesian aggregation, $w_t \propto \exp(-H_t) \cdot \mathrm{ExG}_t$, and the two resolutions are then averaged. A Boltzmann-weighted seasonal phenology prior over the GBIF day-of-year histogram is added in log space at β=1.0, followed by class-prior logit adjustment (τ=0.25) and adaptive probability threshold selection (T=0.03, $k_{\min}$=2, $k_{\max}$=10). The pipeline differs from the anchor (Figure 5) in its multi-scale tiling, ExG screening, and entropy-weighted aggregation; the phenology log-prior is the additional term unique to this submission.

Submission pipeline. Beyond the phenology log-prior itself, the seasonal phenology submission departs from the anchor inference scaffold (Figure 5) in three ways, illustrated architecturally in Figure 12 and visually in Figure 14. First, an Excess Green index $\mathrm{ExG}_t = 2G_t - R_t - B_t$ is computed per tile and tiles with fewer than 15% green pixels are dropped before the encoder forward pass, removing tiles that contain no vegetation. Second, the surviving tiles are aggregated with a per-tile Bayesian weight $w_t \propto \exp(-H_t) \cdot \mathrm{ExG}_t$, where H_t is the predictive entropy of tile t, yielding a softmax-weighted mean rather than the anchor's unweighted mean. Third, the input quadrat is processed at two grid scales (1.0 and 0.8) in addition to the two inference resolutions (224 and 336 pixels), with all per-tile distributions pooled into a single posterior before the phenology prior is added. Logit adjustment (τ=0.25) and adaptive threshold selection (T=0.03) then follow as in the anchor.

Empirical result. On the PlantCLEF 2026 leaderboard, the phenology prior at β=1.0, T=0.03 produced a Macro-F1 of **0.41346** versus the unaugmented BioCLIP 2.5 anchor of 0.41826 ($\Delta = -0.0048$). We swept $\beta \in \{0.25, 0.5, 1.0, 1.5, 2.0\}$ at three probability thresholds; the natural Bayesian operating point (β=1.0, no temperature scaling, T=0.03, τ=0.25 matching the anchor) flips 12.8% of top-one picks at mean k=3.24. The headline result is summarised in §4.7.

Why the gain is negligible. The day-of-year distribution of the test set is heavily summer skewed (Jul = 517, Aug = 668, Apr–Jun = 607 of 2,105 quadrats; Figure 13), so a date-conditioned prior could only help in the densely sampled months. Three concurrent factors plausibly explain the near-zero delta: *(i)* GBIF observation effort is itself summer biased, diluting the discriminative season signal; *(ii)* the BioCLIP 2.5 backbone trained on dated images already encodes seasonal phenotype implicitly via leaf / flower / fruit state; *(iii)* 11% of species fall back to a uniform pdf precisely on the long tail where the visual model is most uncertain, so phenology cannot break ties where it would be most useful.

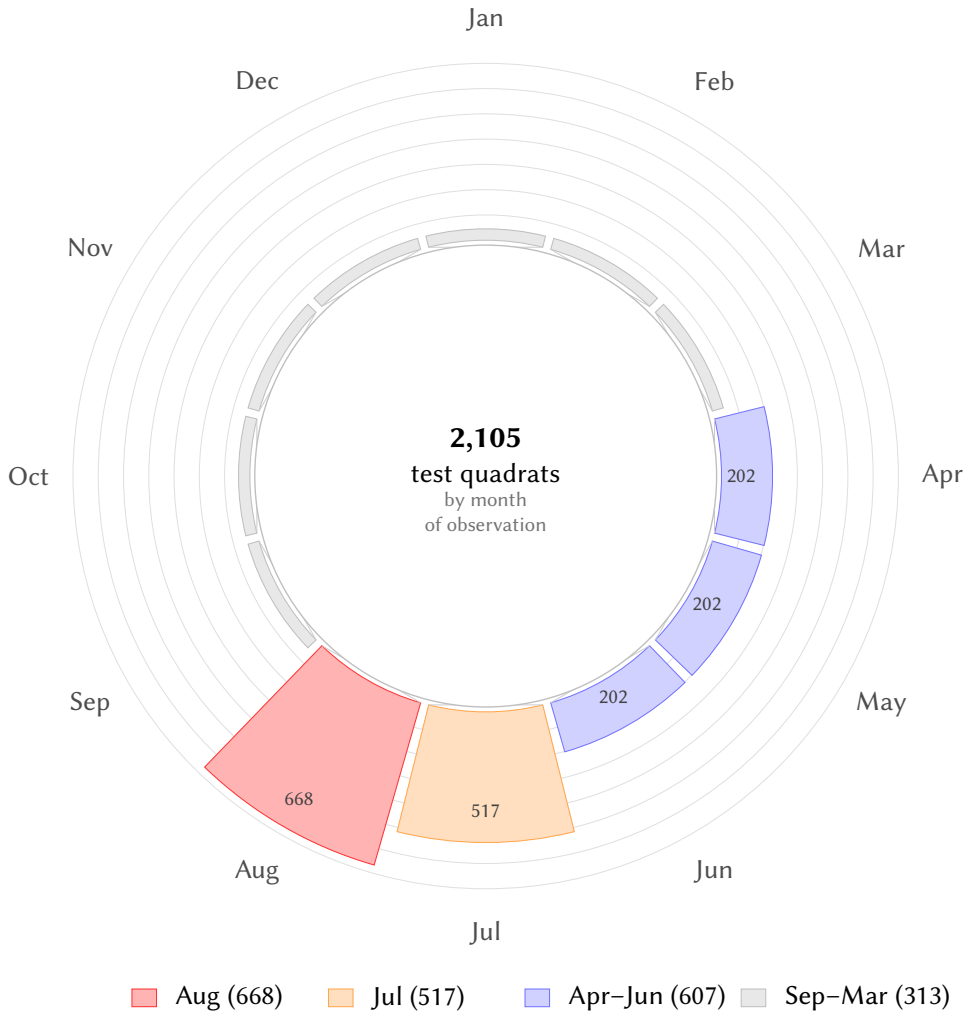

Figure 13: Calendar wheel of the PlantCLEF 2026 test set by month of observation (2,105 quadrats total). Radial bar length is proportional to the count for that month. The July (517) and August (668) wedges are reported exactly in the data; the April–June total of 607 is split uniformly across the three months for the figure, as is the residual 313 across the remaining seven months. The visually obvious summer over-representation motivates the seasonal phenology pivot: a date-conditioned prior could only ever help in the densely-sampled months, and conversely should not be expected to lift Macro F1 on the sparse winter tail.

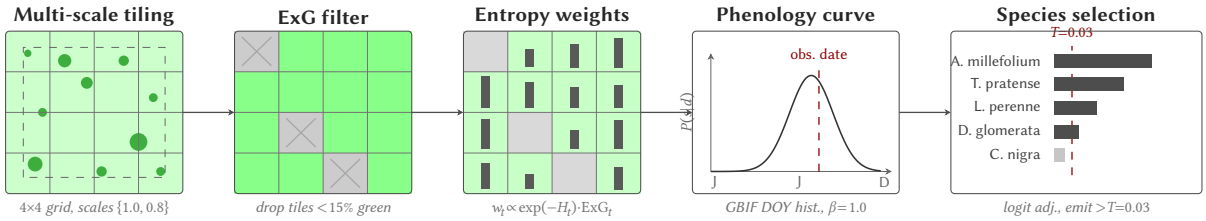

Figure 14: Data flowing through the seasonal phenology submission, stage by stage. *Multi-scale tiling:* the input quadrat is partitioned into a 4×4 grid at the native scale (solid lines) and re-partitioned at scale 0.8 (dashed inner rectangle), giving two tile sets. *ExG filter:* each tile's Excess Green index $\mathrm{ExG}_t = 2G_t - R_t - B_t$ is computed and tiles with fewer than 15% green pixels are dropped (greyed out with a cross). *Entropy weights:* the surviving tiles are weighted by $w_t \propto \exp(-H_t) \cdot \mathrm{ExG}_t$, where bar height encodes the weight; confident, vegetation-rich tiles dominate the softmax-weighted mean. *Phenology curve:* the per-species observability density $P(s \mid d)$, smoothed from the GBIF month-of-observation histogram, is multiplied into the posterior in log space at β=1.0, where the dashed red line marks the quadrat's observation date d. *Species selection:* after class-prior logit adjustment, species whose probability exceeds T=0.03 are emitted; the bottom bar (greyed) falls below threshold and is rejected.

E. Considered Approaches Not Implemented

During the project we scoped a number of additional directions that did not make it into the final system, either because the engineering cost outweighed the expected gain on this benchmark or because preliminary work indicated negative or noise-floor deltas. We list them briefly here for completeness, without method-level detail. Gated multi-backbone fusion with a learned per-image gating network over BioCLIP 2.5, DINOv3, and ConvNeXt-V2-L was prototyped but never completed end-to-end; an offline-cached high-resolution knowledge distillation pipeline with two SWA-averaged frozen teachers was specified but not trained; a graph neural network head over a blended ecological and phylogenetic adjacency matrix was designed and partially implemented; Markov random field reasoning over quadrat tiles via loopy belief propagation was scoped; retrieval-augmented classification using per-species feature prototypes was scoped for the extreme tail; test-time training via self-supervised rotation prediction was scoped; conformal prediction for set-valued species output was scoped; and Frank-Wolfe optimisation with an island-biogeography count prior was scoped as an alternative to the threshold-based selection rule. None of these are evaluated in this paper.

F. Additional Inference Sweeps

This section holds the exhaustive inference-sweep rows that are summarised but not tabulated in §4.6. All numbers are public Macro F1.

Gap and relative-threshold selection rules. Holding τ=0.25 fixed on the i002 last-four-blocks single-resolution checkpoint, the gap rule (emit while consecutive species differ by less than a fraction of the top probability) and the relative-threshold rule (emit every s with p_s within a fraction of the top) both underperform the probability-threshold anchor (0.40590) by between 0.06 and 0.10 Macro F1 (Table 13).

Noisy-or aggregation mixture. On the earlier 010 four-block anchor we additionally swept softmax-mean against a mixture $\alpha \cdot \mathrm{softmax\_mean} + (1-\alpha) \cdot \mathrm{noisy\_or}$ across $\alpha \in \{0.25, 0.50, 0.75, 1.0\}$. All variants land within 0.001 Macro F1 of the softmax-mean anchor at 0.38333, with the second-best, α=0.75, at 0.38278, so on that earlier posterior the choice of aggregator is essentially flat. On the fine-tuned i002 checkpoint, by contrast, softmax-mean pulls clearly ahead of max (§4.5).

Rule	Value	Public Macro F1
gap	0.40	0.33199
gap	0.50	0.33664
gap	0.60	0.34996
relative threshold	0.15	0.31348
relative threshold	0.20	0.31704
relative threshold	0.25	0.31823
relative threshold	0.30	0.33623

Table 13
Gap and relative-threshold selection rules at τ=0.25 on the i002 last-four-blocks single-resolution checkpoint. Both families underperform the probability-threshold anchor (T=0.03, 0.40590; see Table 6) by between 0.06 and 0.10 Macro F1.

Acknowledgments

This work was conducted as part of the COMP3242 Deep Learning course at the Australian National University. We gratefully acknowledge Google's TPU Research Cloud (TRC) program, which provided the Cloud TPU resources that supported part of this research.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools to assist with code development, debugging, and language editing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Picek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Goëau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqué, T. Denton, K. Barnard, C. Nédellec, L. Deléger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of LifeCLEF 2026: Ai challenges for biodiversity understanding and ecosystem management, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2026.

[2] G. Martellucci, I. Moummad, H. Goëau, P. Bonnet, F. Vinatier, A. Joly, Overview of PlantCLEF 2026: Identify multi-species plants in images of vegetation plots, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.

[3] S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong, W. M. Dahdul, C. Stewart, T. Berger-Wolf, W.-L. Chao, Y. Su, BioCLIP: A vision foundation model for the tree of life, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 19412–19422.

[4] J. Gu, S. Stevens, E. G. Campolongo, M. J. Thompson, N. Zhang, J. Wu, A. Kopanev, Z. Mai, A. E. White, J. Balhoff, W. Dahdul, D. Rubenstein, H. Lapp, T. Berger-Wolf, W.-L. Chao, Y. Su, BioCLIP 2: Emergent properties from scaling hierarchical contrastive learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2025. arXiv:2505.23883, spotlight.

[5] H. Goëau, P. Bonnet, A. Joly, The ImageCLEF 2013 plant identification task, in: CLEF 2013 Working Notes, 2013.

[6] S. H. Lee, C. S. Chan, S. J. Mayo, P. Remagnino, How deep learning extracts and learns leaf features for plant classification, Pattern Recognition 71 (2017) 1–13.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Trans-

formers for image recognition at scale, in: International Conference on Learning Representations (ICLR), 2021.

[8] H. Goëau, P. Bonnet, A. Joly, Overview of PlantCLEF 2024: Multi-label identification in vegetation plots, in: CLEF 2024 Working Notes, CEUR Workshop Proceedings, 2024.

[9] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, A. Vedaldi, J. Tolan, P. Labatut, P. Bojanowski, et al., DINOv3, arXiv preprint arXiv:2508.10104 (2025).

[10] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, S. Xie, ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16133–16142.

[11] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 82–91.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.

[13] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, S. Kumar, Long-tail learning via logit adjustment, in: International Conference on Learning Representations (ICLR), 2021.