Vision-Language Guided Plant Localization for Multi-Label Species Prediction in Quadrat Images*

Desmond Khai Tie Oh[1,*], Abel Yu Hao Chai[2] and Sue Han Lee[1]

[1]Swinburne University of Technology Sarawak Campus, 93350, Sarawak, Malaysia
[2]Department of Artificial Intelligence, NEUON AI, 93350, Sarawak, Malaysia

Abstract

This working note presents the approach developed by team **[NEUON AI - Desmond]** for the PlantCLEF multi-species vegetation plot identification task, achieving 20th place on the official leaderboard. The task is challenging because each quadrat image may contain multiple overlapping plant species, while the available classifier is mainly trained on single-plant images. To address this issue, we investigate a two-stage inference framework that combines local Vision-Language Model spatial reasoning with DINOv2-based species classification. In the first stage, a local VLM is used as a weak localization module to identify visually distinct plant groups through structured grid-based prompts. Instead of asking the VLM to predict exact scientific species names, we use it to group plant regions with similar visual features and return their spatial locations in a machine-readable format. In the second stage, the selected regions are cropped and passed to the organizer-provided DINOv2 classifier for species-level prediction. We further evaluate the effects of SAM3 segmentation, 3×3 and 8×8 tiling, JPEG recompression and interpolation-based preprocessing, test-time augmentation, aggregation strategies, and different local VLM models. Experimental results show that VLM-guided localization can improve multi-species prediction when combined with localized DINOv2 inference. The best-performing configuration uses gemma4-based localization, 8×8 + 3×3 tiling, JPEG recompression and interpolation, test-time augmentation, and MAX aggregation. Our primary submission incorporating SAM3 segmentation achieved a peak public score of 0.35643, while the non-segmented baseline demonstrated superior robustness on the private evaluation.

Keywords

PlantCLEF, multi-species plant identification, vegetation quadrat images, Vision-Language Model, DINOv2, spatial localization, test-time augmentation, SAM3 segmentation

1. Introduction

Automatic identification of plant species from vegetation plot images is an important task for biodiversity monitoring, ecological surveys, and long-term environmental assessment. Traditional botanical surveys usually require expert knowledge and manual inspection in the field, which is time-consuming when large numbers of plots must be analyzed. The PlantCLEF challenge addresses this problem by encouraging the development of computer vision systems capable of predicting all plant species visible in high-resolution quadrat images. Unlike conventional single-label plant identification, where an image usually contains one dominant plant or a close-up organ, the vegetation plot setting requires multi-label recognition from complex natural scenes containing many overlapping species, different growth stages, shadows, background objects, and partially visible plant structures.

From PlantCLEF 2024 until PlantCLEF 2026, the tasks focus on multi-species identification in vegetation quadrats. In this setting, the input image is a high-resolution photograph of a small vegetation plot, and the system must output the list of plant species present in the scene. The task is especially difficult because of the strong domain shift between the available training data and the test data. The training set mainly consists of single-label images of individual plants or plant organs, while the test set contains dense multi-species vegetation plots captured in real field conditions. Previous PlantCLEF overview reports indicated that the training data contains approximately 1.4 million single-plant images covering 7,806 species, whereas the test set contains 2,105 high-resolution quadrat images from Pyrenean, Mediterranean, and South-Western European temperate floras [1, 2, 3]. These test images

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
123desmondoh@gmail.com (D. K. T. Oh); abel_cyh@neuon.ai (A. Y. H. Chai); shlee@swinburne.edu.my (S. H. Lee)

Figure 1: Sample vegetation plots with wooden frames, measuring tapes, shadows, blur, soil, dead material, and small plant parts.

may include wooden frames, measuring tapes, shadows, blur, soil, dead material, and small plant parts as shown in Fig. 1, making direct whole-image classification unreliable.

The organizers provided DINOv2-based Vision Transformer models pre-trained and fine-tuned on PlantCLEF single-plant data to support participants [4]. These models offer strong visual representations and a practical starting point for inference without requiring participants to train large-scale models from scratch. However, directly applying a single-plant classifier to an entire vegetation plot is limited because the model may focus on dominant plants, background texture, or irrelevant objects, while missing small or partially occluded species. Previous approaches therefore commonly adopted patch-wise or tile-based inference, where each high-resolution plot is divided into smaller crops before classification [5, 6, 7, 8, 9]. For example, DS@GT used a 4×4 tiling strategy with a PlantCLEF-fine-tuned Vision Transformer and visual-cluster priors, showing that tile-level inference can better match the model's 518×518 receptive field and improve multi-species prediction [6].

In this work, we investigate how far Vision-Language Models (VLM) can contribute to the PlantCLEF vegetation plot identification task. Most traditional vision-only models, including the organizer-provided DINOv2 classifier, process images mainly as visual patterns and produce class probabilities based on visual similarity. Although these models can be highly effective for plant recognition, they do not naturally express the image content in language or reason explicitly about semantic descriptions unless they are specifically designed or trained with language supervision. This limitation is important in vegetation plot images because the scene often contains multiple overlapping species, visually distinct plant groups, dead material, background objects, and small plant structures that are difficult to handle using whole-image classification alone.

VLM provide a different opportunity. Instead of only producing numerical class scores, they can describe and organize visual content through language. In our setting, this means the model can explain the visual structure of a quadrat image by separating visible plant groups, describing them as anonymous visual categories, and assigning them to spatial regions. We therefore explore whether a local VLM can act as a weak semantic and spatial reasoning module for the challenge. The goal is not to rely on the VLM to directly identify scientific plant names, since fine-grained botanical classification remains difficult and the VLM may hallucinate species. Instead, we use the VLM to answer a simpler but useful question: where are the visually distinct plant groups located in the vegetation plot?

To investigate this idea, we conduct a series of experiments combining preprocessing, segmentation, Vision-Language Model reasoning, and DINOv2-based species inference. We explore how different combinations of preprocessing strategies, segmentation approaches, spatial decomposition methods, and inference aggregation techniques affect vegetation plot recognition performance. Our team, **[NEUON AI - Desmond]**, secured the 20^{th} position in the final PlantCLEF 2026 leaderboard evaluation.

flora and vegetation, including plants, grasses, leaves, leafs, flowers, tree, bark, tree branches, and pile of dead leaves on the ground

Figure 2: The prompt used for SAM3 Segmentation.

2. PlantCLEF Overview

Recent PlantCLEF approaches demonstrated that post-processing and inference design can be as important as model training. The Webmaking solution optimized a frozen DINOv2 model using multi-scale tiling, test-time augmentation, artifact down-weighting with zero-shot segmentation, and ecological priors such as GBIF occurrence statistics, seasonal plausibility, and species co-occurrence [8]. Another approach explored zero-shot segmentation through prototype guidance, using class prototypes from the training set to guide attention toward plant-relevant regions before classification [9]. A separate working note emphasized that preprocessing alone can significantly affect performance, especially choices such as interpolation method, JPEG recompression quality, and chroma subsampling after resizing [5]. These works suggest that the main difficulty is not only recognizing plant species, but also deciding where and how to apply the classifier inside a cluttered quadrat image.

These methods indicate that the challenge is not only a classification problem, but also a localization and scene understanding problem. High-resolution vegetation plots often contain overlapping species, partially visible plant organs, dead material, and visually distracting background structures. As a result, determining where and how to apply species classifiers becomes an important component of the pipeline. Motivated by these observations, we investigate whether Vision-Language Models can contribute semantic and spatial reasoning abilities to assist vegetation plot understanding.

3. Methodology

3.1. Problem Formulation

The PlantCLEF vegetation plot recognition task requires identifying multiple plant species from a single plot-level image. Unlike single-object classification, each image may contain several visually overlapping plant species, where different plants may appear at different spatial locations, scales, and levels of visibility. This makes direct whole-image classification challenging because the classifier must simultaneously localize relevant plant regions and infer species identity from fine-grained visual cues.

In this work, we formulate the task as a two-stage inference problem. Instead of directly predicting species from the entire vegetation plot image, we first identify spatial regions that are likely to contain visually distinct plant groups, and then perform species classification on these localized regions. This formulation separates spatial reasoning from fine-grained species recognition, allowing each component to focus on a different part of the multi-species recognition problem.

3.2. Segmentation with SAM3

Before applying the two-stage inference pipeline, we investigate whether segmentation can improve the visibility of plant regions by removing irrelevant background information. For this purpose, we use the SAM3 segmentation model with text prompts to generate plant-focused masks from the original vegetation plot images [10].

The text prompt is designed to guide the segmentation model toward living plant regions while avoiding irrelevant background elements such as soil, stones, and shadows. The prompt used in our experiment is shown in Fig. 2.

The segmented outputs are then compared with the original images to observe whether the segmentation process improves plant visibility or removes useful species-level details. Fig. 3 shows examples of

Figure 3: Comparison between the original vegetation plot image and the SAM3 segmentation result.

the original image and the corresponding SAM3-segmented result.

3.3. Two-stage Inference Framework

We propose a two-stage inference framework that combines Vision-Language Model spatial reasoning with DINOv2-based species classification. The main idea is to use a local Vision-Language Model as a weak localization module to identify where visually distinct plant groups appear in the image, before applying a species classifier to the localized regions.

In the first stage, the Vision-Language Model analyzes the vegetation plot image using structured grid-based prompts and returns the spatial locations of visually distinct plant groups. In the second stage, the DINOv2 classifier performs species inference on the cropped image regions generated from the first stage. The predictions from multiple localized regions are then aggregated to produce the final multi-species prediction for the test image.

3.3.1. Stage 1: VLM-guided Plant Region Localization

The first stage, named *VLM-guided Plant Region Localization*, uses a local Vision-Language Model (VLM) to identify spatial regions containing visually distinct plant groups. This is because the VLM may not have sufficiently fine-grained botanical recognition ability to accurately predict exact scientific species names. Therefore, instead of asking the VLM to perform species-level identification, we use it to group plant regions with similar visual features, such as leaf shape, growth form, color, texture, and density. The species-level identification is then handled by the DINOv2 classifier in the second stage.

The VLM receives an image together with a structured prompt that defines a grid coordinate system. The model is instructed to ignore irrelevant regions such as black background, soil, dry grass, dead stems, shadows, stones, and uncertain fragments. It then returns the grid cells that contain each visually distinct living plant group in a strict JSON format.

Grid-based Tiling. To support spatial reasoning, each input image is divided into grid-based regions. We experiment with both 3×3 and 8×8 grid decompositions, as shown in Fig. 4. The 3×3 grid provides coarse spatial localization, while the 8×8 grid provides finer region-level localization.

The grid coordinates returned by the VLM are mapped back to the corresponding image regions. These selected regions are then cropped from the original image and passed to the second-stage

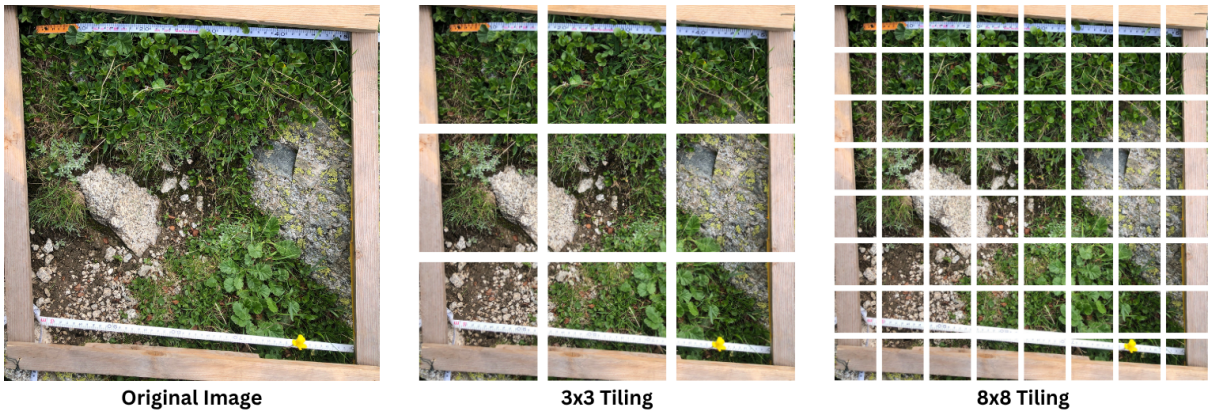

Figure 4: Example of 3×3 and 8×8 grid overlays used for VLM-guided plant region localization. The highlighted cells indicate regions selected by the VLM for second-stage classification.

classifier. This allows the model to focus on localized plant regions instead of using only the full vegetation plot image. The grid overlay also makes it possible to visualize which regions were selected by the VLM and how they correspond to the cropped image patches used for classification.

Prompt Engineering for VLM Localization. We design the VLM prompt to produce structured and machine-readable localization outputs. Since the output from Stage 1 is later used to generate cropped images for Stage 2 classification, the response format must be consistent and easy to parse. Therefore, instead of asking the VLM to freely describe the image, we guide it to return plant group locations using predefined grid cell names.

The prompt explicitly defines the grid coordinate system, the expected JSON output format, and several filtering rules. These rules instruct the VLM to ignore irrelevant regions such as black background, soil, stones, shadows, dry grass, dead stems, brown debris, and uncertain fragments. The VLM is also instructed to group plants based on visual appearance rather than scientific species names. This is important because Stage 1 is designed only for localization, while species identification is handled by the DINOv2 classifier in Stage 2. An example of the structured prompt used for VLM localization as shown in Fig. 5.

For the finer localization setting, the same prompt structure is adapted to an 8×8 grid by redefining the grid cells from R1C1 to R8C8. This allows the VLM to provide more detailed spatial locations, which are then converted into localized image crops for DINOv2-based species classification.

3.3.2. Stage 2: DINOv2-based Local Species Classification

The second stage, named *DINOv2-based Local Species Classification*, performs species prediction on the localized image regions obtained from Stage 1. We use the DINOv2-based classifier provided by the PlantCLEF organizers as the main species recognition model. Instead of applying the classifier only to the whole image, we apply it to the VLM-selected local regions so that fine-grained plant details can be evaluated more directly.

For each test image, the selected grid cells are converted into image crops. We then apply data preprocessing to these crops before passing them into the DINOv2 classifier. The purpose of this preprocessing step is to make the test images as close as possible to the expected input distribution of the DINOv2-based model. The preprocessed crops are then passed into the DINOv2 classifier, producing species-level prediction scores for each localized region. We also apply Test-Time Augmentation (TTA) by generating multiple augmented versions of each crop and aggregating their prediction scores. This allows the final prediction to be more stable across different image transformations.

Data Preprocessing and Augmentation. To standardize the input to the DINOv2 classifier and

```
Identify distinct visible plant groups in the image and assign each group to one
or more positions.

Important:
- You do not need to know the real scientific species name.
- Use "species 1", "species 2", etc. only as visual group labels.
- If plants are visible, do not return {} just because you are unsure of the
  exact species identity.
- Return {} only when there are no visible plants or the image content is
  impossible to inspect.

Return ONLY valid JSON.
Do not explain.
Do not include markdown.
Do not include any text before or after the JSON.

Use this format:
{
  "species 1": ["R1C1", "R1C2"],
  "species 2": ["R3C3"]
}

Row = R
Column = C

Allowed positions only:
R1C1, R1C2, R1C3,
R2C1, R2C2, R2C3,
R3C1, R3C2, R3C3,
whole image

Rules:
- Each key must be like "species 1", "species 2", "species 3", and so on.
- Each value must be a list of allowed positions only.
- If only one plant species appears across the whole image, return:
  {"species 1": ["whole image"]}
- If no plant is visible, return:
  {}
```

Figure 5: The text prompt used for local VLM during Stage 1: VLM-guided Plant Region Localization.

mitigate domain discrepancies arising from image compression or resolution variances, we adopt the preprocessing workflow from [5], applying JPEG recompression and interpolation-based resizing. Furthermore, Test-Time Augmentation (TTA) is employed during inference by generating multiple augmented versions of each localized crop. The resulting region-level prediction scores are subsequently aggregated to yield a more stabilized and robust species prediction.

3.4. Inference Procedure

The complete inference process starts from a test vegetation plot image, as illustrated in Fig. 6. The proposed pipeline is divided into two main stages: VLM-guided plant region localization and DINOv2-based local species classification.

Before localization, the original vegetation plot image can optionally be processed using SAM3 segmentation. This step is used to investigate whether removing background regions, such as soil, shadows, and non-plant areas, can make living plant regions more visible for the following localization stage. The segmented or original image is then passed into Stage 1.

In Stage 1, a local Vision-Language Model is used to identify the spatial positions of visually distinct plant groups. Instead of asking the VLM to directly predict scientific species names, the model is

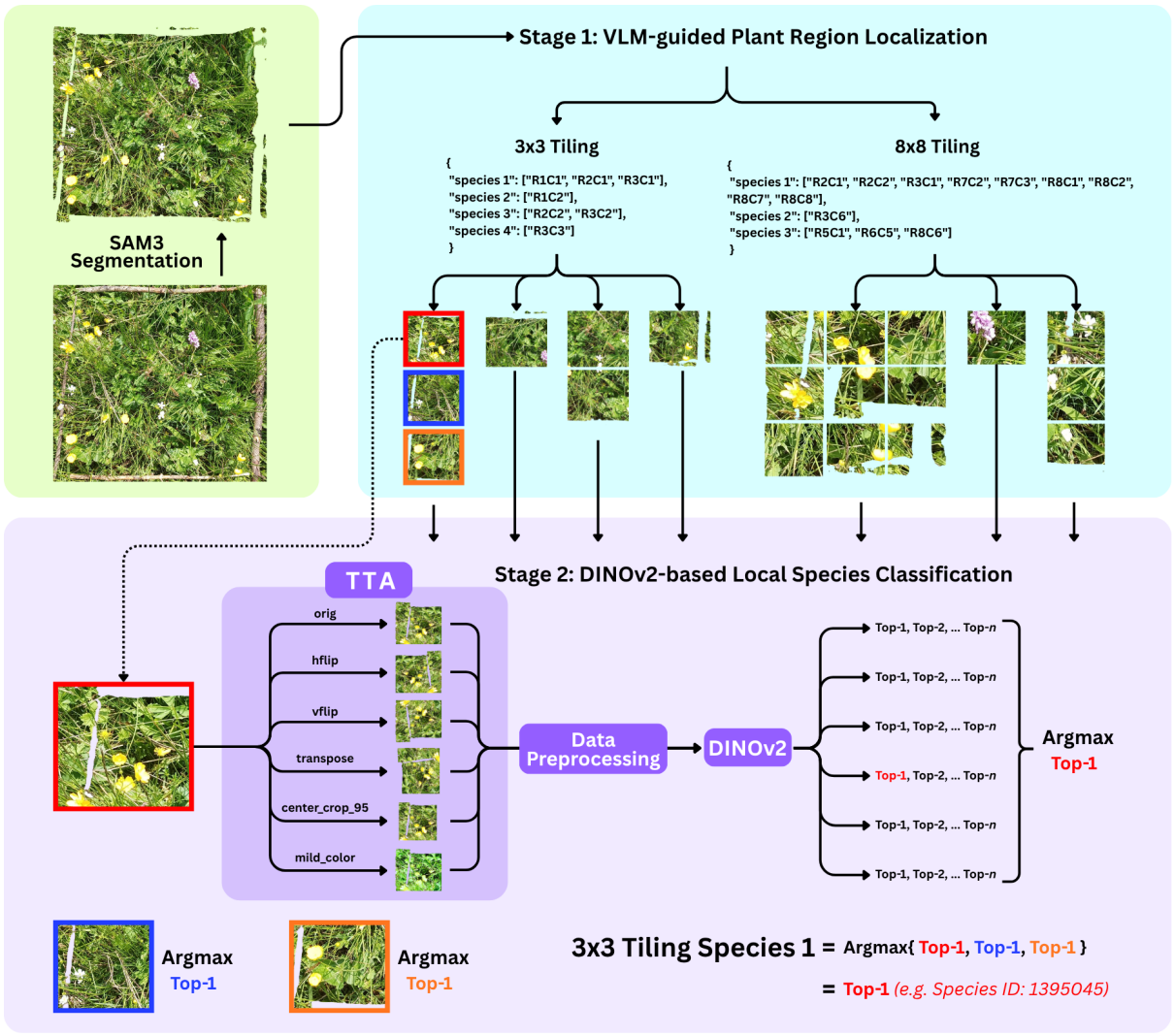

Figure 6: Proposed two-stage inference framework. The vegetation plot image is first segmented using SAM3, followed by VLM-guided plant region localization with 3×3 and 8×8 grid tiling. The localized regions are cropped and passed to a DINOv2-based classifier for local species classification, before aggregating the predictions into the final multi-species output.

prompted to locate plant regions using grid-based reasoning. Two localization settings are explored: 3x3 tiling and 8x8 tiling. For each setting, the image is divided into grid cells, and the VLM returns a structured output indicating which cells contain each visually distinct plant group. For example, in the 3x3 setting, a plant group may be localized to cells such as R1C1, R2C1, and R3C1, while the 8x8 setting provides a finer spatial localization of smaller plant regions.

The selected grid cells are then converted into image crops. In the 3x3 setting, larger and more general plant regions are extracted, while the 8x8 setting produces smaller and more precise crops. These crops are treated as candidate local regions that may contain one or more target species. As shown in the figure, each selected region is independently forwarded to the second stage for classification.

In Stage 2, each localized crop is classified using the provided DINOv2-based species classifier. Before inference, each crop is processed using the same preprocessing strategy, including JPEG recompression and interpolation-based resizing. Test-time augmentation is also applied to each crop, including transformations such as the horizontal flip, vertical flip, transpose, center crop, and mild colour adjustment. The DINOv2 model produces a ranked list of candidate species for each augmented crop, such as Top-1, Top-2, until Top-n predictions.

The predictions from different augmentations are then aggregated to obtain the final prediction

for each localized region. For each crop, the most confident species prediction is selected using an argmax operation over the aggregated prediction scores. Finally, predictions from all selected regions are combined to form the final multi-species prediction for the original vegetation plot image.

Overall, the proposed inference pipeline separates the multi-species plant recognition problem into two complementary components. The VLM is responsible for language-guided spatial localization, where it identifies visually distinct plant regions without performing scientific species classification. The DINOv2 classifier is then responsible for local species recognition on the selected crops. This design allows the pipeline to investigate how far a local VLM can contribute to the PlantCLEF multi-species recognition task by providing spatial reasoning, while relying on a stronger vision-based classifier for fine-grained species identification.

4. Experiments and Results

4.1. Experimental Overview

We conduct a series of experiments to investigate how Vision-Language Models can assist vegetation plot understanding when combined with the organizer-provided DINOv2 classifier. Instead of introducing a single fixed architecture, we evaluate multiple combinations of preprocessing strategies, tiling configurations, segmentation approaches, aggregation methods, and VLM-based spatial decomposition techniques.

The experiments are designed to evaluate six main factors in the proposed inference pipeline: the number of predicted species, aggregation strategies, the effect of test-time augmentation, comparison of local VLM models, the effect of tiling size, and the effect of SAM3 segmentation.

4.2. Evaluation Settings and Metrics

To evaluate the performance of multi-species identification, we strictly follow the official PlantCLEF 2026 competition protocol. The primary evaluation metric is the **transect-averaged macro F_1-score per sample**. This metric balances precision and recall for each multi-label plot image independently.

To mitigate spatial sampling bias from over-sampled areas, the test set is structured into distinct *transects* (e.g., 5m $\times$ 1m surveyed areas). The final leaderboard score is computed by first calculating the macro-averaged F_1-score within each transect i, and then averaging across all N transects:

$$\text{Final Score} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{|T_i|} \sum_{j \in T_i} F_1(y_j, \hat{y}_j) \right) \tag{1}$$

where N is the total number of transects, T_i is the number of quadrats in transect i, and $F_1(y_j, \hat{y}_j)$ is the sample-level F_1-score for the j-th quadrat image based on True Positives (TP_j), False Positives (FP_j), and False Negatives (FN_j). Our team submitted under the official name **[NEUON AI - Desmond]**, ultimately achieving the **20th place** on the final official leaderboard.

4.3. Number of Predicted Species

The PlantCLEF vegetation plot task requires predicting multiple plant species within a single quadrat image. Previous PlantCLEF overview analysis reported that the average number of species per quadrat is approximately 8 [3]. Therefore, we first investigate different strategies for determining the final number of predicted species during inference.

Two main approaches were explored. The first approach always outputs a fixed number of species predictions based on the average species count observed in the dataset. The second approach allows the Vision-Language Model to estimate the number of visually distinct plant groups present in the quadrat image, and the final number of predicted species follows the VLM estimation. These experiments

Table 1
Evaluation of predicted species count strategies on the PlantCLEF 2026 leaderboard (Metric: Transect-Averaged Macro F_1, Higher is Better).

VLM Model	Tiling	JPEG + Interp.	TTA	Agg.	Remark	Private F_1	Public F_1
qwen3-vl	3×3	False	False	SUM	8 species	0.2229	0.2280
qwen3-vl	3×3	False	False	SUM	VLM-selected regions	**0.25938**	**0.26964**

were conducted to evaluate whether adaptive species count estimation can better match the complex composition of real vegetation plots.

The results in Table 1 demonstrate that allowing the local VLM to adaptively guide the number of selected crops based on visual distinctness significantly outperforms forcing a static prediction constraint. Following the VLM-selected regions yields a clear improvement, raising the Public F_1 from 0.22800 to 0.26964 and Private F_1 from 0.22290 to 0.25938. Using a fixed baseline of 8 species introduces distinct limitations. Although the average species density across the training plots is approximately 8, real-world quadrats vary drastically. Imposing a strict cutoff forces the downstream classifier to introduce false positives in species-sparse plots (degrading precision) or prematurely truncate true species in highly biodiverse plots (degrading recall). Similar to the aggregation trends, during the active phase of the competition, this dynamic strategy was consistently validated via the public leaderboard feedback (representing approximately 11% of the test data). It firmly established that language-guided spatial decomposition accommodates the irregular botanical density inherent in multi-label vegetation monitoring far better than global dataset statistics.

4.4. Aggregation Strategy

After the VLM identifies plant regions and the DINOv2 classifier generates predictions for each localized crop, the regional predictions need to be combined into a final image-level species ranking. We compare two aggregation strategies: SUM aggregation and MAX aggregation.

SUM aggregation accumulates prediction scores across all localized regions, which favors species that appear consistently across multiple crops. MAX aggregation selects the highest prediction score for each species from any localized region, which favors species that are strongly detected in at least one crop. This comparison helps us understand whether repeated evidence or strongest regional evidence is more useful for vegetation plot species prediction.

Table 2
Comparison of SUM and MAX aggregation strategies (Metric: Transect-Averaged Macro F_1). † denotes the primary configuration selected based on public leaderboard feedback.

VLM Model	Tiling	JPEG + Interp.	TTA	Agg.	Remark	Private F_1	Public F_1
qwen3-vl	$8 \times 8 + 3 \times 3$	True	True	SUM	Score accumulation across regions	**0.34492**	0.29526
qwen3-vl	$8 \times 8 + 3 \times 3$	True	True	MAX	Highest regional selection †	0.33008	**0.32694**

The results show different behavior between private and public scores. SUM aggregation achieves the highest private score, suggesting that accumulating evidence from multiple localized regions can improve robustness when the same species appears in several crops. However, MAX aggregation obtains a higher public score, indicating that the strongest local evidence may be more reliable for some images, especially when plant species are only visible in small or specific regions. During the active phase of the competition, model selection was strictly guided by the public leaderboard feedback (representing approximately 11% of the test data), where MAX aggregation yielded superior performance. However, post-competition evaluation on the private leaderboard (89% of the test data) revealed that

SUM aggregation provides better robustness, indicating that accumulating continuous regional evidence generalizes better to unseen test transects.

4.5. Effect of Test-time Augmentation

Test-time augmentation is applied during DINOv2 inference to reduce sensitivity to image transformation, crop position, and visual variation. In this experiment, we compare inference with and without TTA while keeping the VLM model, tiling configuration, preprocessing strategy, and aggregation method fixed.

The TTA setting includes multiple transformed views of the same localized crop, including the original image, horizontal flip, vertical flip, transpose, center crop, and mild color adjustment. The prediction scores from these augmented views are combined to obtain a more stable regional prediction.

Table 3
Effect of test-time augmentation (TTA) during DINOv2 inference (Metric: Transect-Averaged Macro F_1).

VLM Model	Tiling	JPEG + Interp.	TTA	Agg.	Remark	Private F_1	Public F_1
qwen3-vl	$8 \times 8 + 3 \times 3$	True	False	MAX	Without test-time augmentation	0.30755	0.31613
qwen3-vl	$8 \times 8 + 3 \times 3$	True	True	MAX	With test-time augmentation	**0.33008**	**0.32694**

The results indicate that TTA improves both private and public scores. This suggests that the DINOv2 classifier benefits from observing multiple transformed versions of the same localized plant region. Since plant crops may contain different orientations, partial organs, shadows, and scale variations, TTA helps stabilize the classifier output and reduces dependence on a single fixed image view.

4.6. Comparison of Local VLM Models

We also compare different local Vision-Language Models for Stage 1 plant region localization. Since the VLM is not used for direct species classification, its main role is to provide useful spatial decomposition by identifying visually distinct plant groups. Therefore, a better VLM should produce region selections that are more helpful for the downstream DINOv2 classifier.

In this experiment, qwen3-vl and gemma4 are evaluated using the same tiling strategy, preprocessing setting, TTA setting, and aggregation method.

Table 4
Performance comparison of different local Vision-Language Models for Stage 1 localization (Metric: Transect-Averaged Macro F_1).

VLM Model	Tiling	JPEG + Interp.	TTA	Agg.	No. of Parameters	Private F_1	Public F_1
qwen3-vl	$8 \times 8 + 3 \times 3$	True	True	MAX	4b	0.33008	0.32694
gemma4	$8 \times 8 + 3 \times 3$	True	True	MAX	31b	**0.34542**	**0.33784**

The gemma4 model achieves higher private and public scores than qwen3-vl under the same experimental setting. This indicates that the quality of VLM-guided spatial localization has a direct influence on the final species prediction performance. Although the VLM does not output scientific species names, better region localization can provide more informative crops to the DINOv2 classifier, improving the final multi-species ranking.

4.7. Effect of Tiling Size

The tiling strategy controls how the vegetation plot image is spatially decomposed before VLM localization. A coarse grid, such as 3×3, provides larger regions and preserves more context, while a fine grid, such as 8×8, provides more precise localization but may produce smaller crops with less surrounding information.

We compare the combination of 8×8 and 3×3 tiling against an additional setting that also includes the full image. The purpose is to test whether adding the full image can improve species prediction by preserving global context.

Table 5
Effect of different tiling configurations on the leaderboard (Metric: Transect-Averaged Macro F_1).

VLM Model	Tiling	JPEG + Interp.	TTA	Agg.	Remark	Private F_1	Public F_1
gemma4	$8 \times 8 + 3 \times 3$	True	True	MAX	Cropped images only	**0.34542**	**0.33784**
gemma4	$8 \times 8 + 3 \times 3 + 1$	True	True	MAX	Cropped + full image	0.22041	0.2115

The result shows that adding the full image significantly decreases performance. This suggests that whole-image inference may introduce noisy background information and many visually irrelevant regions, which weakens the benefit of VLM-guided localization. In contrast, using only the selected 8×8 and 3×3 regional crops allows the classifier to focus more directly on plant-rich areas. This supports the design choice of separating spatial localization from species classification.

4.8. Effect of SAM3 Segmentation

We further evaluate whether SAM3 segmentation can improve the inference pipeline by removing background noise before VLM localization and DINOv2 classification. The goal of segmentation is to highlight living plant regions and reduce the influence of irrelevant visual elements such as soil, stones, dry grass, shadows, and dead plant debris.

Two segmentation settings are evaluated. The first uses SAM3 segmentation with one concise text prompt, while the second uses a fuller segmentation setting. Both are compared against the non-segmented baseline using the same VLM model, tiling configuration, preprocessing strategy, TTA setting, and aggregation method.

Table 6
Ablation study on SAM3-based pre-segmentation versus grid-based tiling (Metric: Transect-Averaged Macro F_1). ⋆ indicates the final primary submission of our team.

VLM Model	Tiling	JPEG + Interp.	TTA	Agg.	Remark	Private F_1	Public F_1
gemma4	$8 \times 8 + 3 \times 3$	True	True	MAX	Without SAM3 segmentation	**0.34542**	0.33784
gemma4	$8 \times 8 + 3 \times 3$	True	True	MAX	SAM3 with one prompt ⋆	0.32625	**0.35643**
gemma4	$8 \times 8 + 3 \times 3$	True	True	MAX	SAM3 full setting	0.33102	0.35053

As outlined in Table 6, the configuration incorporating SAM3 with one concise prompt achieved our peak public leaderboard score of **0.35643**. Because the public leaderboard was the only visible feedback loop during the active phase of the competition (representing approximately 11% of the test data), this specific pipeline was logically selected as the official primary submission for team **[NEUON AI - Desmond]**. However, post-competition unblinding on the private leaderboard (the remaining 89% of the test data) revealed a minor shake-down, where the baseline without SAM3 proved more robust

(0.34542). This indicates that while SAM3 successfully isolates precise visual contours on the public split to maximize localized precision, it may occasionally over-segment or discard faint, overlapping plant structures across the broader private test distribution.

5. Overall Findings

Across the experiments, several findings can be observed. First, allowing the VLM to guide the number of selected regions performs better than forcing the prediction list to contain a fixed number of species. This suggests that adaptive visual decomposition is more suitable for vegetation plot images than a fixed output assumption.

Second, regional inference is more effective than adding the full image into the prediction pipeline. The large performance drop observed when using 8×8 + 3×3 + full-image inference indicates that global image predictions may introduce unnecessary background noise and reduce the advantage of localized classification.

Third, TTA consistently improves performance, showing that the DINOv2 classifier benefits from multiple augmented views of localized plant crops. Fourth, the choice of local VLM is important. gemma4 provides better downstream results than qwen3-vl under the same setting, suggesting that more accurate spatial reasoning can improve final species classification even without direct species prediction by the VLM.

Finally, SAM3 segmentation shows potential but is not consistently better across private and public scores. Segmentation can help suppress background noise, but it may also remove small or fine-grained plant structures that are important for species recognition. Overall, the best-performing strategy combines gemma4-based VLM localization, 8×8 + 3×3 tiling, JPEG recompression and interpolation, TTA, and MAX aggregation.

Acknowledgments

We gratefully acknowledge the support provided through the Inspiring Women in Science: Scientific Achievement Award 2025 by Springer Nature. We also gratefully acknowledged the support of NEUON AI with the GPU workstation used for this research. We also thank the PlantCLEF organizers for providing the challenge dataset, pretrained DINOv2 model, and evaluation platform.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-5.5 for grammar checking, language refinement, and improving the clarity of the writing. All AI-generated suggestions were reviewed, edited, and verified by the author(s). No Generative AI system was used to create original scientific ideas, conduct experiments, analyze data, or draw conclusions. The author(s) take full responsibility for the content of this publication.

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, 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] G. Martellucci, H. Goeau, P. Bonnet, F. Vinatier, A. Joly, Overview of plantclef 2025: Multi-species plant identification in vegetation quadrat images, 2025. URL: https://arxiv.org/abs/2509.17602. arXiv:2509.17602.

[4] H. Goëau, J.-C. Lombardo, A. Affouard, V. Espitalier, P. Bonnet, A. Joly, Plantclef 2024 pretrained models on the flora of the south western europe based on a subset of pl@ntnet collaborative images and a vit base patch 14 dinov2, 2024. URL: https://doi.org/10.5281/zenodo.10848263. doi:10.5281/zenodo.10848263.

[5] V. Espitalier, Preprocessing is all you need: Theheartofnoise submission to plantclef 2025, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025. URL: https://ceur-ws.org/Vol-4038/paper_238.pdf.

[6] M. Gustineli, A. Miyaguchi, A. Cheung, D. Khattak, Tile-based vit inference with visual-cluster priors for zero-shot multi-species plant identification, 2025. URL: https://arxiv.org/abs/2507.06093. arXiv:2507.06093.

[7] H. Herasimchyk, R. Labryga, T. Prusina, Multi-label plant species prediction with metadata-enhanced multi-head vision transformers, 2025. URL: https://arxiv.org/abs/2508.10457. arXiv:2508.10457.

[8] N. Semenova, Optimizing a vision transformer with ecological context for multi-label plant species identification, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 3166–3178. URL: https://ceur-ws.org/Vol-4038/paper_392.pdf.

[9] L. A. D. Filho, A. M. da Silva Neto, R. P. David, R. T. Calumby, Zero-shot segmentation through prototype-guidance for multi-label plant species identification, 2025. URL: https://arxiv.org/abs/2512.19957. arXiv:2512.19957.

[10] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollár, N. Ravi, K. Saenko, P. Zhang, C. Feichtenhofer, Sam 3: Segment anything with concepts, 2026. URL: https://arxiv.org/abs/2511.16719. arXiv:2511.16719.