Prompt Sensitivity and Fine-Tuning of CLIP for Medicinal Plant Recognition

No Author Given

No Institute Given

Abstract. Identifying medicinal plants automatically remains a difficult problem due to the scarcity of annotated images, large visual variations within species, and the diversity of naming conventions across languages. Recent vision–language models such as CLIP make it possible to perform classification without task-specific training by using natural language prompts, thereby reducing dependence on extensive datasets. Yet, their applicability to biodiversity-related domains and multilingual contexts has not been adequately examined. In this study, we assess the performance of CLIP for the recognition of Ayurvedic medicinal plants using a dataset collected at the Panchakarma Research Centre in Trivandrum, India. We design experiments on 2-, 10-, and 20-class subsets to test three aspects: (i) the influence of prompt phrasing (short versus descriptive text), (ii) the effect of language choice (English compared with transliterated Malayalam), and (iii) the value of lightweight fine-tuning strategies. Our results indicate that while zero-shot CLIP performs well on simple binary distinctions, its accuracy decreases as the number of classes increases. Detailed prompts in English improve zero-shot accuracy, but the highest gains are obtained through frozen-encoder fine-tuning with a linear classifier, achieving 81.82 % on 10 classes and 70.21 % on 20 classes. Transliteration-based prompts were less reliable, underscoring the need for multilingual and culturally aware adaptations of vision–language systems.

Keywords: CLIP, Vision–Language Models, Medicinal Plant Recognition, Prompt Engineering, Fine-Tuning, Multilingual Computing

1 Introduction

Ayurveda, a traditional Indian system of medicine, uses medicinal plants for everything from digestive issues to long-term chronic conditions. Correctly telling one plant from another is not just a technical task—it matters for safe medical use, conservation of species, and for checking whether old remedies stand up to modern scientific study. In practice, though, it is not easy. Even trained eyes can struggle to separate similar leaves when conditions change, such as poor light or overlapping branches in the field. Building such expertise takes years, and not everyone has access to it. This is where digital methods have started to play a role. Computer vision offers a practical shortcut, though current deep learning

systems usually assume the availability of very large, carefully labeled image collections. For most medicinal plants, especially regional ones, such datasets hardly exist.

The earliest computer-based efforts at plant recognition leaned on manually designed features—things like the outline of a leaf, the patterns of its veins, or the texture on its surface [1, 2]. These descriptors worked fairly well when the photographs were taken in controlled conditions, but they often failed once plants were observed outdoors, where light changes, cluttered backgrounds, and partial occlusions are the norm. The arrival of deep learning marked a turning point. Convolutional networks such as ResNet [3] and EfficientNet [4] demonstrated strong results on large image benchmarks, and later studies adapted them to datasets like PlantCLEF [5]. Even so, the promise of these models comes with a heavy demand: thousands of labeled samples per species. For rare or locally specific plants, assembling that kind of dataset is usually unrealistic.

Researchers have also tried to ease the burden of large-scale annotation through few-shot and attention-based methods. For instance, Jiang et al. [7] and Zhu et al. [8] showed that attention mechanisms and meta-learning could classify plant diseases with relatively small datasets. Other attempts have turned to cross-lingual augmentation to widen biodiversity coverage, such as the work of Deng et al. [9] and Banerjee et al. [10]. While these efforts helped reduce reliance on exhaustive labeling, their scope was still narrow. In practice, models trained this way often remained tied to specific tasks and did not generalize well to new domains.

A more transformative direction came with the rise of vision–language models (VLMs). CLIP [11], trained on a massive collection of 400 million image–text pairs, demonstrated that visual and linguistic information could be embedded in a shared space, enabling zero-shot transfer to unseen tasks. This idea quickly grew into larger multimodal frameworks such as ALIGN [12] and Florence [13]. At the same time, efficient adaptation strategies were introduced, including CLIP-Adapter [14], CoOp [15], and Tip-Adapter [17], which reduced the computational overhead of domain-specific fine-tuning. The agricultural domain has begun adopting these models as well—examples include AgriCLIP [20] for crop imagery, VLCD [21] for plant disease detection, and BioTrove-CLIP [22] for large-scale biodiversity classification.

Despite these advances, little work has examined how multilingual prompts or transliteration affect performance, particularly in biodiversity contexts like Ayurvedic medicinal plants. Since CLIP is heavily English-centric, prompt phrasing and language choice can significantly influence results. Moreover, while prompt tuning has been proposed as a lightweight alternative to retraining, its efficacy in specialized plant recognition remains underexplored.

In this paper, we systematically evaluate CLIP for medicinal plant identification. Our experiments investigate (i) prompt phrasing (short vs. descriptive), (ii) language effects (English vs. transliterated Malayalam), and (iii) fine-tuning strategies (classifier-head vs. prompt fine-tuning). Section II reviews related work on plant recognition and VLMs. Section III describes our dataset collection from the Panchakarma Ayurvedic Research Centre and outlines the methodology. Section IV presents results across 2-, 10-, and 20-class scenarios, while Section V analyzes error patterns and qualitative outcomes. Section VI provides discussion and benchmarks against related approaches, and Section VII concludes with key findings and directions for future research.

2 Related Work

2.1 Traditional and Deep Learning for Plant Recognition

The earliest computer-based efforts at plant recognition leaned on manually designed features—things like the outline of a leaf, the patterns of its veins, or the texture on its surface [1, 2]. These descriptors worked fairly well when the photographs were taken in controlled conditions, but they often failed once plants were observed outdoors, where light changes, cluttered backgrounds, and partial occlusions are the norm. The arrival of deep learning marked a turning point. Convolutional networks such as ResNet [3] and EfficientNet [4] demonstrated strong results on large image benchmarks, and later studies adapted them to datasets like PlantCLEF [5]. Even so, the promise of these models comes with a heavy demand: thousands of labeled samples per species. It's usually not practical to create that kind of dataset for rare or local plants.

2.2 Attention, Few-Shot, and Cross-Lingual Methods

To mitigate data scarcity, attention-based methods and few-shot learning gained traction. Jiang et al. [7] leveraged attention mechanisms for disease classification in leaves, while Zhu et al. [8] applied meta-learning to handle small training sets. Cross-lingual augmentation was introduced to address regional language challenges [9], and Banerjee et al. [10] curated Indian medicinal plant datasets. These studies reduced the need for large-scale annotation but still needed specialized adjustments, which made it hard to apply them to different tasks.

2.3 Vision–Language Models and Adaptation

The advent of large-scale vision–language pretraining changed the field. CLIP [11] demonstrated that pairing images and text embeddings enabled zero-shot classification across domains. ALIGN [12] and Florence [13] further scaled multimodal

datasets and architectures. To adapt CLIP efficiently, Gao et al. [14] proposed CLIP-Adapter, Zhou et al. [15] introduced CoOp with learnable context vectors, and Zhou et al. [16] extended this with CoCoOp for unseen classes. Zhang et al. [17] developed Tip-Adapter, a training-free cache mechanism. Other methods, such as semantic-aware fine-tuning [18] and visual prompt tuning [19], showed ways to make adaptations with fewer parameters.

2.4 Applications in Agriculture and Biodiversity

Recent studies adapted VLMs for agriculture and biodiversity. Nawaz et al. [20] introduced AgriCLIP, integrating agricultural image–text data. Zhou et al. [21] presented VLCD for crop leaf disease classification using descriptive prompts. Li et al. [22] released BioTrove-CLIP, covering 33k species with multimodal supervision. Chen et al. [23] proposed RegionCLIP, and Zhang et al. [24] designed RAFTer, both leveraging region attention for fine-grained recognition. More recently, multilingual adaptations such as M-CLIP [25] demonstrated the potential of extending CLIP beyond English. Even with these advancements, there's been little research focused on how multilingual prompts perform in biodiversity fields, which is an important gap that our work aims to address.

3 Dataset and Methodology

3.1 Data Collection

To build a domain-specific benchmark, we collected a curated dataset from the **Panchakarma Ayurvedic Research Centre, Poojapura, Trivandrum**. Each species class contains approximately 220–250 high-quality images. The collection focused on widely used Ayurvedic medicinal plants such as *Tulsi* (*Ocimum tenuiflorum*), *Amla* (*Phyllanthus emblica*), *Neem* (*Azadirachta indica*), and *Curry leaves* (*Murraya koenigii*). Images were captured under both natural and semi-controlled conditions, with variations in lighting, background clutter, and partial occlusions to reflect realistic field conditions. All images were reviewed manually to ensure quality, and duplicates or heavily corrupted samples were discarded.

3.2 Dataset Splits

To systematically evaluate scaling behavior, three experimental subsets were defined:

- **2-class subset:** a binary task, used to probe the lower bound of zero-shot performance.
- **10-class subset:** a mid-sized challenge covering representative medicinal plants with higher intra-class variability.

– **20-class subset:** a fine-grained classification task designed to test CLIP's robustness to closely related species.

Each subset was split into 70% training, 15% validation, and 15% testing, ensuring that images from the same sample were not shared across splits.

3.3 Preprocessing

All images were resized to 224×224 pixels to match CLIP's ViT-B/32 input specification. Standard normalization was applied using CLIP's mean and variance values. Data augmentation was kept minimal—random horizontal flips and slight rotations—since strong augmentations risk distorting biologically important features such as venation and leaf shape.

3.4 Workflow Overview

Figure 1 illustrates the overall pipeline. The dataset first undergoes preprocessing, after which prompts are designed in English and transliterated Malayalam (short and detailed). The CLIP model is then applied in two modes: (i) zero-shot classification using only prompts, and (ii) fine-tuning using classifier-head or prompt-tuning approaches. Finally, predictions are evaluated using accuracy and confusion matrices.

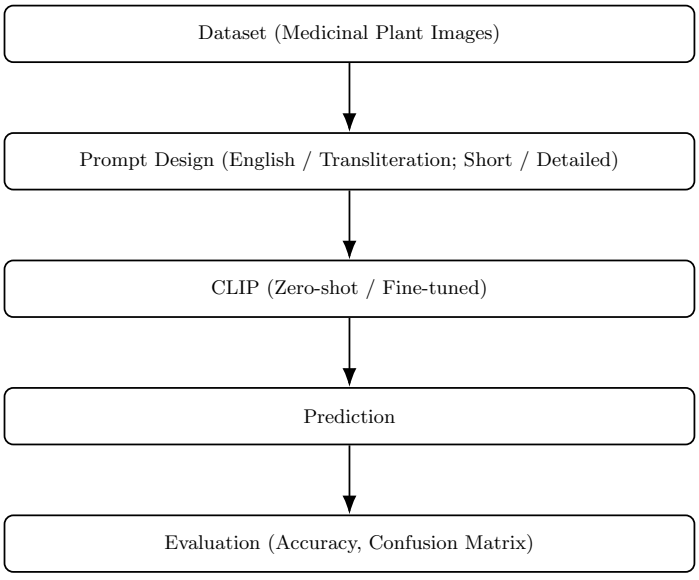

Fig. 1. Workflow: dataset → prompt design → CLIP (zero-shot / fine-tuned) → prediction → evaluation.

3.5 Fine-Tuning Strategies

Two parameter-efficient fine-tuning strategies were compared:

1. **Classifier-Head Fine-Tuning:** the CLIP image encoder was frozen, and a linear classifier head was trained on top of the extracted embeddings. This method is computationally light and adapts decision boundaries without retraining the encoder [14].

2. **Prompt Fine-Tuning:** learnable context vectors were prepended to class names, following CoOp [15] and CoCoOp [16]. While promising in prior work, our experiments showed that prompt tuning alone underperformed relative to classifier-head fine-tuning, consistent with observations in Tip-Adapter [17].

Given this disparity, we emphasize classifier-head fine-tuning in our main results.

3.6 Training Setup

Training employed the AdamW optimizer with learning rate 5×10^{-4}, weight decay 10^{-4}, and batch size of 16. Early stopping was applied with a patience of 5 epochs based on validation loss. For zero-shot experiments, no training was performed; instead, predictions were derived directly from prompt embeddings.

3.7 Mathematical Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the dataset of N leaf images x_i with class labels $y_i \in \{1, 2, \ldots, C\}$, where C is the number of classes.

CLIP Embeddings CLIP encodes an image x and a text prompt t into a shared embedding space [11]:

$$v = f_\theta(x), \quad u = g_\phi(t), \tag{1}$$

where f_θ is the image encoder and g_ϕ is the text encoder. Both embeddings are ℓ_2-normalized:

$$\hat{v} = \frac{v}{\|v\|}, \quad \hat{u} = \frac{u}{\|u\|}. \tag{2}$$

Zero-Shot Classification The similarity between embeddings is measured with cosine similarity, scaled by a temperature parameter τ [11]. The probability of class c with text description t_c is:

$$P(y = c|x) = \frac{\exp(\tau \cdot \hat{v}^\top \hat{u}_c)}{\sum_{k=1}^{C} \exp(\tau \cdot \hat{v}^\top \hat{u}_k)}. \tag{3}$$

The predicted class is $\hat{y} = \arg\max_c P(y = c|x)$.

Table 1. Accuracy (%) across prompt styles and fine-tuning on the 10- and 20-class subsets. A dash (–) indicates not evaluated in that configuration.

Setting	10 Classes		20 Classes	
	Zero-shot	Fine-tuned	Zero-shot	Fine-tuned
English (short)	29.09	81.82	52.00	70.21
English (detailed)	47.27	–	56.00	–
Transliterated (short)	10.91	74.55	08.00	75.53
Transliterated (detailed)	25.45	–	59.00	–

*(*i) Zero-shot uses frozen CLIP with prompts only;
(ii) Fine-tuned uses a frozen encoder with a trained linear head;
(iii) Results are top-1 accuracy on held-out test splits.

Classifier-Head Fine-Tuning In frozen-encoder fine-tuning [14,15], the image encoder is fixed, and a linear classifier $W \in \mathbb{R}^{C\times d}$ is trained on embeddings v:

$$z = Wv, \quad P(y = c|x) = \frac{\exp(z_c)}{\sum_{k=1}^{C} \exp(z_k)}. \tag{4}$$

The training objective is the cross-entropy loss [3]:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log P(y_i|x_i). \tag{5}$$

Prompt Fine-Tuning Prompt tuning introduces learnable context vectors $\{p_j\}_{j=1}^{M}$ concatenated with the class name t_c [15, 16]. The resulting text embedding is:

$$u_c = g_\phi([p_1, p_2, \ldots, p_M, t_c]). \tag{6}$$

In practice, this strategy achieved modest improvements in our dataset but consistently lagged behind classifier-head fine-tuning, echoing limitations reported in Tip-Adapter [17].

4 Results

Table 1 summarizes quantitative performance across zero-shot and fine-tuned CLIP configurations. Visual comparison is provided via learning curves (Fig. 2 and Fig. 3).

4.1 Binary (2-Class) Evaluation

The simplest scenario tested CLIP's ability to distinguish between two visually distinct classes. Zero-shot CLIP achieved **100%** accuracy across both English and transliterated prompts, regardless of prompt style. This suggests that when

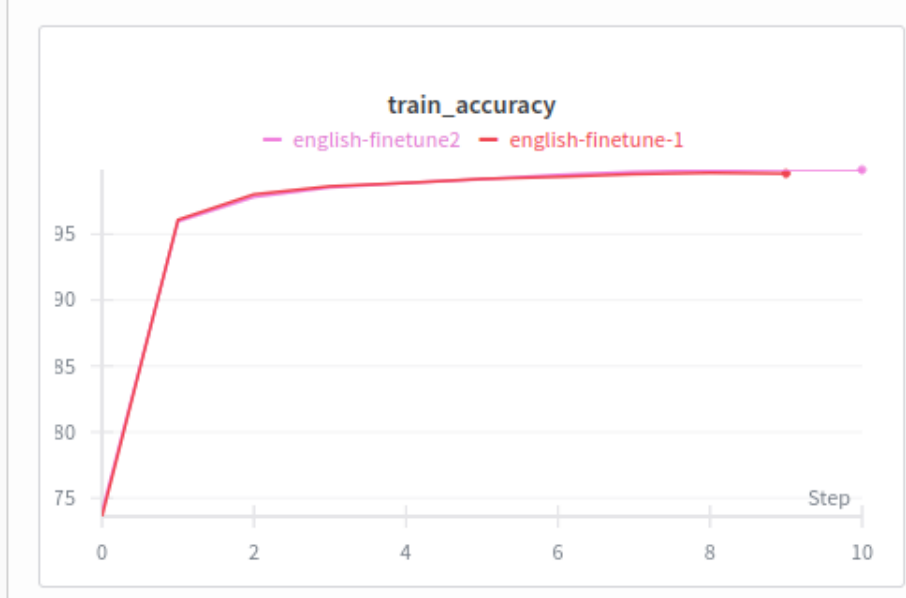

Fig. 2. Training accuracy

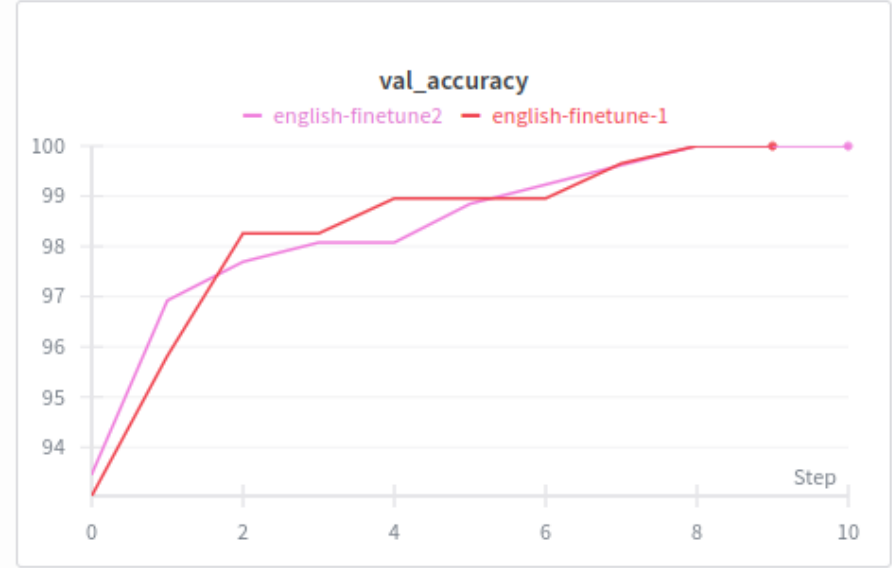

Fig. 3. Test accuracy

the decision space is small, CLIP's pretrained embeddings are sufficiently separable without adaptation. While encouraging for rapid deployment, the binary task under-represents real-world complexity in medicinal plant identification.

4.2 Ten-Class Evaluation

The 10-class setting introduces higher intra-class similarity and cross-class confusion. Zero-shot CLIP with **short English prompts** achieved only **29.09%**. Incorporating descriptive attributes improved zero-shot performance to **47.27%**, demonstrating sensitivity to prompt richness. Transliteration performed poorly (**10.91**–**25.45%**), indicating that CLIP's text encoder does not effectively map transliterated Malayalam tokens to visual concepts.

Classifier-head fine-tuning delivered a substantial gain, reaching **81.82%** for English and **74.55%** for transliterated Malayalam, far exceeding zero-shot baselines. The learning curves in Fig. 2 and Fig. 3 shows steady convergence without overfitting, confirming that a small supervised head on frozen features can adapt effectively. Prompt fine-tuning plateaued near 70% in our trials and did not match the classifier-head approach, reinforcing our decision to emphasize the latter.

4.3 Twenty-Class Evaluation

Scaling to 20 classes further stresses zero-shot generalization. English prompts yielded **52**–**56%** accuracy, while transliteration produced mixed outcomes (**8**–**59%**). In a few cases, transliterated tokens partially aligned with CLIP's subword statistics, but performance remained unstable overall.

Classifier-head fine-tuning again proved robust, achieving **70.21%** for English and **75.53%** for transliterated Malayalam. Although lower than the 10-class result (as expected with increased granularity), the gain over zero-shot remains substantial. The learning showed stable convergence. Misclassifications were more distributed across fine-grained confusable pairs, that emphasize global contours over venation and surface textures.

4.4 Quantitative Comparisons

Figure 4 visualizes performance differences across configurations, reinforcing three trends:

- **Prompt richness helps** zero-shot CLIP, but descriptive gains diminish as class cardinality increases.
- **English prompts outperform transliteration**, reflecting CLIP's English-centric pretraining corpus.
- **Classifier-head fine-tuning dominates**, closing the gap to supervised performance with minimal compute.

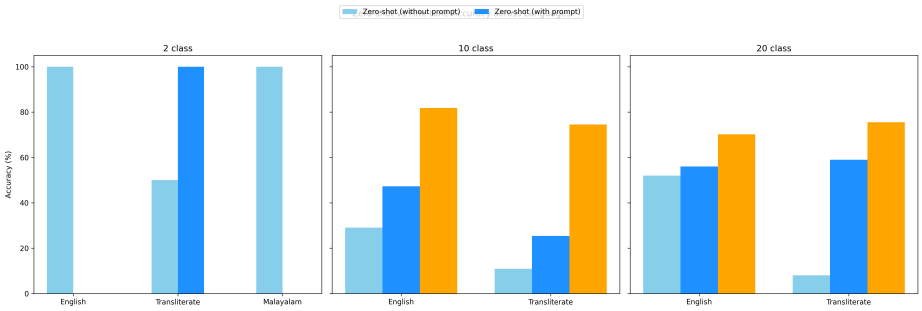

Fig. 4. Performance Visualization

4.5 Error Analysis and Qualitative Insights

Error patterns suggest that pretraining priors strongly influence zero-shot predictions. Confusions among near-neighbor species emphasize the need for fine-grained cues. Region-aware adaptations (e.g., RegionCLIP, RAFTer) and adapter-based tuning may improve sensitivity to venation and localized textures, reducing systematic errors.

4.6 Summary of Findings

Overall, the results show that: (i) zero-shot CLIP suffices for trivial binary distinctions but underperforms for multi-class recognition; (ii) prompt engineering improves zero-shot but cannot overcome domain shift; (iii) frozen-encoder *classifier-head fine-tuning* consistently yields the best accuracy and scales more gracefully to larger label spaces; and (iv) transliteration introduces instability and should be used cautiously until multilingual pretraining is strengthened.

5 Discussion

5.1 Multilinguality and Prompt Sensitivity

Our results demonstrate that CLIP's performance is highly sensitive to prompt phrasing and language. Detailed prompts consistently outperform short ones, confirming that richer textual context improves alignment with CLIP's training distribution. However, transliterated Malayalam prompts underperformed relative to English, with accuracy sometimes dropping by more than 20 percentage points in the 10-class scenario. This is consistent with the fact that CLIP's text encoder was primarily trained on English web text [11]. While transliteration occasionally matched CLIP's subword embeddings and achieved competitive results in the 20-class setting, performance was unstable. These findings align with broader work showing that multilingual or cross-lingual augmentation is needed to generalize beyond English [9, 25]. Multilingual adaptations such as

M-CLIP [25] have already demonstrated the feasibility of extending CLIP to non-English corpora, and similar strategies may be vital for biodiversity domains where species names are culturally and linguistically diverse.

5.2 Fine-Tuning Efficiency

A second major finding is the effectiveness of frozen-encoder fine-tuning with a classifier head. This approach consistently delivered strong gains, improving accuracy from 47.27% zero-shot (English detailed prompts, 10 classes) to 81.82% with minimal computational cost. In contrast, prompt fine-tuning plateaued near 70%, failing to match classifier-head results. This suggests that adaptation on the text side alone may be insufficient when domain shift is primarily visual. Our observations echo the conclusions of CoOp [15], CLIP-Adapter [14], and Tip-Adapter [17], where light-weight feature adaptation strategies outperform naive prompt manipulation. For biodiversity applications, where large-scale retraining is impractical, frozen-encoder approaches provide an attractive trade-off between efficiency and accuracy.

5.3 Future Works

The error analysis highlights that CLIP often confuses morphologically similar classes (e.g., *Tulsi* vs. *Mint*), indicating that global embeddings are insufficient for fine-grained recognition. Region-aware adaptations, such as RegionCLIP [23] and RAFTer [24], which incorporate localized attention, represent promising future directions. Another open challenge is mitigating pretraining bias. Domain-specific adaptations such as AgriCLIP [20] or BioTrove-CLIP [22] suggest that curated, domain-aligned datasets can reduce such biases. Finally, multilingual extensions should be explored, building on works like M-CLIP [25] to ensure robustness when applying CLIP in culturally diverse biodiversity contexts.

6 Conclusion

This paper presented a systematic study of applying CLIP to medicinal plant recognition, with a particular focus on prompt sensitivity, transliteration, and parameter-efficient fine-tuning strategies. Using a curated dataset collected from the Panchakarma Ayurvedic Research Centre in Trivandrum, we evaluated CLIP across 2-, 10-, and 20-class subsets. Our experiments revealed three key findings. First, zero-shot CLIP is sufficient for trivial binary distinctions but struggles as the class space expands. Second, prompt phrasing and language substantially affect performance: detailed English prompts improved accuracy over short ones, while transliterated Malayalam prompts produced unstable results. Third, frozen-encoder fine-tuning with a classifier head consistently delivered the highest gains, surpassing both zero-shot and prompt-tuning approaches, and scaling

robustly to 20 classes.

Beyond quantitative metrics, our analysis uncovered error patterns tied to pretraining biases, and qualitative evidence that CLIP attends primarily to global contours rather than fine venation. These insights underline the need for region-sensitive adaptation and multilingual extensions. Compared to prior biodiversity and agricultural studies [20–22], our work uniquely highlights the interaction between prompt design, transliteration, and lightweight fine-tuning in an Ayurvedic context.

This work contributes new evidence that while CLIP offers a promising foundation for biodiversity informatics, careful prompt engineering alone is insufficient in specialized domains. Lightweight fine-tuning emerges as an effective and practical strategy. Looking ahead, future research should integrate multilingual encoders, region-aware attention modules, and curated biodiversity datasets to further enhance robustness. By doing so, vision–language models may become powerful tools for supporting traditional medicine, conservation, and healthcare applications in linguistically and biologically diverse settings.

References

1. S. Lee, "Leafsnap: Mobile plant recognition using computer vision," 2018.
2. J. Barbedo, "Plant species recognition from photographs: a review," *Computers and Electronics in Agriculture*, 2018.
3. K. He, et al., "Deep residual learning for image recognition," *CVPR*, 2019.
4. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling," *ICML*, 2019.
5. S. Girod, et al., "The PlantCLEF benchmark," *CLEF*, 2019.
6. Y. Wu, "Plant recognition with deep CNNs," *Pattern Recognition*, 2020.
7. X. Jiang, "Leaf disease classification using attention networks," *Sensors*, 2020.
8. H. Zhu, "Few-shot learning for plant disease detection," *ACM MM*, 2020.
9. X. Deng, "Cross-lingual data augmentation for biodiversity," *ACL*, 2020.
10. A. Banerjee, "Indian medicinal plant datasets," *IJCAI Workshop*, 2020.
11. A. Radford, et al., "Learning transferable visual models from natural language supervision," *ICML*, 2021.
12. C. Jia, et al., "Scaling up vision-language pretraining with ALIGN," *ICML*, 2021.
13. L. Yuan, et al., "Florence: A new foundation model for vision-language," *CVPR*, 2021.
14. C. Gao, et al., "CLIP-Adapter: Better vision-language models with feature adapters," *arXiv preprint*, 2021.
15. K. Zhou, et al., "Learning to prompt for vision-language models," *CVPR*, 2022.
16. K. Zhou, et al., "Conditional prompt learning for vision-language models," *NeurIPS*, 2022.
17. R. Zhang, et al., "Tip-Adapter: Training-free adapter for vision-language models," *ECCV*, 2022.
18. Z. Sun, "Semantic-aware fine-tuning for CLIP," *arXiv*, 2022.
19. Y. Jia, "Visual prompt tuning for vision-language models," *NeurIPS*, 2022.
20. U. Nawaz, et al., "AgriCLIP: Adapting CLIP for agriculture," *arXiv*, 2022.

21. Y. Zhou, et al., "VLCD: Vision-language crop disease detection," *Sensors*, 2023.
22. X. Li, et al., "BioTrove-CLIP: Biodiversity vision-language models," *arXiv*, 2023.
23. S. Chen, et al., "RegionCLIP: Region-based language-vision pretraining," *CVPR*, 2023.
24. H. Zhang, et al., "RAFTer: Region attention fine-tuning for CLIP," *ICLR*, 2023.
25. T. Huang, et al., "M-CLIP: Multilingual contrastive learning," *ACL*, 2023.