

Bridging the Domain Gap in Agricultural Vision: Parameter-Efficient VLM Adaptation via Expert Descriptions

Deeksha Aggarwal Yash Mittal Uttam Kumar
Spatial Computing Laboratory, IIIT Bangalore
Bangalore, India.

{deeksha.aggarwal003, YashMittal006, uttam}@iiitb.ac.in

Abstract

The zero-shot and few-shot capabilities of current vision-language models (VLMs) including CLIP is constrained by the availability of large-scale, aligned image-text paired datasets in agricultural domain. In this work, we leverage two complementary sources of information (i) category-level text descriptions generated by large language models (LLMs) and (ii) open source fine-grained image classification datasets to improve the zero-shot and few-shot classification performance of VLMs in agricultural domain. We propose a resource-efficient framework that leverages LLMs to synthesize exhaustive, class-specific textual descriptions of phenological and morphological traits. These descriptions are used to fine-tune CLIP via Low-Rank Adaptation (LoRA) and a novel category-level contrastive loss that accommodates multiple semantic descriptors per training batch. Our framework significantly outperformed vanilla CLIP and standard CLIP-LoRA baselines in zero-shot and few-shot settings. Critically, it demonstrates superior domain generalization on real-world field datasets, effectively bridging the performance gap between controlled laboratory environments and real world agriculture fields. Our findings suggest that category level detailed text descriptions are effective and are complementary to visual appearance.

1. Introduction

The increasing global demand for food security and resource efficiency has positioned precision agriculture as a critical frontier for technological innovation. Computer vision powered by Deep Learning (DL), has shown immense promise in automating key agricultural monitoring tasks such as plant disease detection and crop classification. These tasks are essential for timely intervention, minimizing crop losses, and increasing crop yield [1]. While DL models trained on large, labeled datasets have achieved

near-human performance in controlled settings, their real-world utility remains constrained [19].

Transitioning computer vision models from the laboratory to real-world agricultural environments reveal significant hurdles due to domain shift and the scarcity of labeled samples from field conditions [5]. A model trained on a controlled source domain typically suffers a dramatic performance drop when deployed on a heterogeneous target domain. The unparalleled variability of agricultural environments exacerbates this shift, with major discrepancies arising across: (i) geographies and climates (e.g., soil type, light intensity), (ii) crop and variety types, (iii) seasons and growth stages (phenology), and (iv) data acquisition methods (e.g., sensor types, drone vs. ground-level imagery). This heterogeneity routinely violates the fundamental assumption of supervised learning—that training and testing data are drawn from an independent and identically distributed (i.i.d.) distribution [4]. Traditional Transfer Learning (TL) techniques, which involve unfreezing and fine-tuning a few layers of a Deep Neural Network (DNN), often fail to achieve the robust generalization required for scalable deployment due to the vast distributional gap and limited labeled target data [2].

Recent breakthroughs in zero- and few-shot classification have been driven by Vision-Language Models (VLMs) like CLIP [13]. These models utilize massive image-text datasets to learn a shared embedding space between visual and natural language domains. However, the performance of these VLMs often falters in highly specialized domains (e.g., healthcare [17], geo-sensing [16], and agriculture [11]). In agriculture, these failures stem from two interrelated issues: (i) intrinsic domain gap: VLMs are pre-trained on general-purpose images of everyday objects, lacking the fine-grained, domain-specific examples necessary to identify subtle plant disease symptoms or detect minor variations among crop species, and (ii) specialized data scarcity: adapting VLMs is hindered by lack of comprehensive labeled data sources from real-world fields. Existing open source and large scale datasets, such as the Plant

Village Dataset (PVD) are laboratory-generated and focus on narrow tasks with only images and class names, restricting their utility for complex VLM fine-tuning. In addition to the difficulty and high resource demands of gathering large-scale, high-quality image-caption datasets from real-world fields, traditional VLM fine-tuning methods present further obstacles. These conventional approaches typically require substantial computational power and extensive labeled datasets to achieve convergence [23], rendering them impractical for many localized agricultural applications where data and hardware are often limited.

Clearly, previous reported research highlight the need for advanced techniques, particularly Parameter-Efficient Fine-Tuning (PEFT), which is crucial when computational resources or labeled data are limited. To address the challenges of domain gaps, data scarcity and limited generalization, we introduce a novel framework that leverages the strengths of Large Language Models (LLMs) and PEFT to adapt VLMs for fine-grained agricultural tasks. We utilize the fact that LLMs can generate rich, class-specific text descriptions of fine-grained categories. These detailed descriptions can be paired with existing vision-based datasets, such as PVD, to generate coarsely-aligned image-text datasets for fine-tuning VLMs on agricultural tasks. This approach is demonstrated to improve zero-shot and few-shot performance, generalizing to new classes and geographical fields. Our method addresses the high computational costs and data limitations that currently restrict foundation models to laboratory settings, providing a scalable alternative for real-time, on-site diagnostics.

Motivated by these challenges, this work provides the following contributions:

- (i) Fine-grained task and class specific textual description generation: We generated class-specific, fine-grained image-text datasets derived solely from existing vision-based agricultural datasets by leveraging targeted LLM prompting.
- (ii) Novel fine-tuning pipeline: We introduce a robust training pipeline that stochastically pairs images with class-category texts, coupled with a novel category-level contrastive loss designed to handle multiple positive texts per image.
- (iii) Parameter-efficient adaptation of CLIP: We employed Low-Rank Adaptation (LoRA) to fine-tune CLIP under resource-constrained settings, significantly reducing computational overhead while ensuring effective knowledge transfer.
- (iv) Real-world field evaluation: The framework is rigorously evaluated using a self collected real-world field dataset to validate its practical generalizability under zero shot settings.

2. Related work

2.1. Vision language models and parameter-efficient fine-tuning

Recently, Vision Language Models (VLMs) [7, 12, 13, 22] have rapidly gained prominence, finding widespread application in both zero-shot and few-shot classification tasks. These models are designed to process image and text data jointly, comprising both specialized image and text encoders, along with fusion mechanisms to learn a shared, semantic embedding space between the two modalities. Following extensive pre-training on massive datasets, VLMs have demonstrated significant capability improvements across various downstream visual tasks. For example, CLIP [13] utilizes an image-text contrastive objective, which quantifies the similarity between corresponding image and text embeddings via a dot product. This contrastive learning mechanism achieves strong alignment of image and text features, enabling impressive zero-shot prediction capabilities when paired with an appropriate text such as the class name during test time.

Researchers have developed a variety of fine-tuning strategies to adapt Vision-Language Models like CLIP for task-specific requirements and to facilitate robust domain transfer. One prominent approach is prompt tuning, where CoOp [25] improves few-shot classification by optimizing learnable prompt context vectors that are appended to class name embeddings. To further enhance generalization, CoCoOp [26] introduced dynamic conditional prompting, while KgCoOp [20] sought to minimize prompt variance to preserve foundational textual knowledge when addressing previously unseen categories. Beyond prompt-based methods, adapter-based tuning introduces small, task-specific trainable modules between pre-trained layers, which effectively reduces the computational overhead and training time required for fine-tuning. More recently, CLIP-LoRA [23] utilized the Low-Rank Adaptation (LoRA) [6] technique to perform parameter-efficient fine-tuning, traditionally evaluating performance using standard template prompts such as “a photo of a [class].”

2.2. Agricultural domain specific vision language models

The drive toward artificial intelligence (AI) solutions tailored for agriculture is evident in the development of domain-specific models. Recent works like AgroGPT [3] and AgriVLM [21] operate primarily as visual question answering (VQA) systems. They rely on the training of LLMs (e.g., the 7B LLaMA model [15] or ChatGLM [24]) on extensive datasets of question-answer pairs or multi-turn conversations derived from image descriptions generated by generic LLMs and external knowledge. Similarly, the CDIP-ChatGLM3 [18] approach uses a dual-model inte-

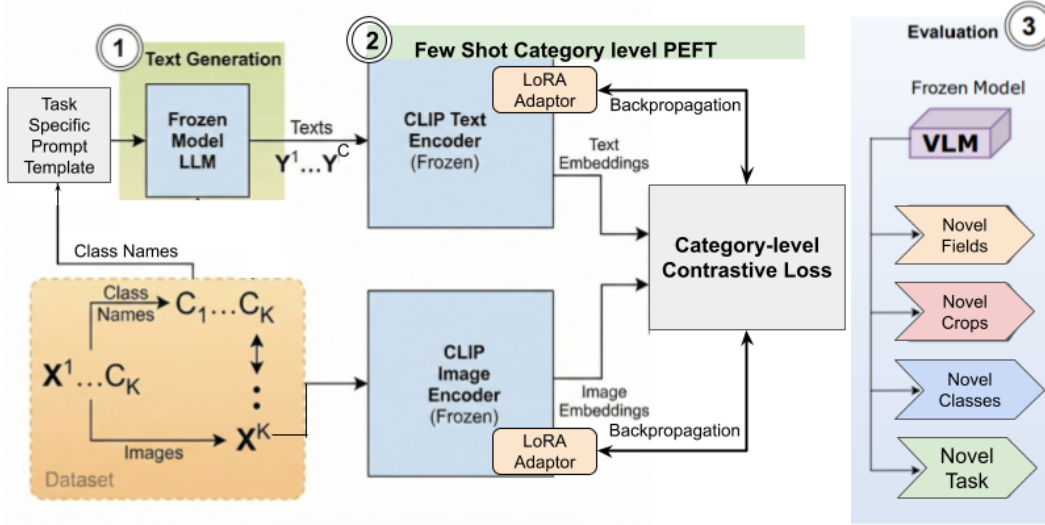


Figure 1. Proposed framework demonstrating LoRA based parameter efficient fine-tuning (PEFT) of CLIP using category level contrastive loss. 1) Text generation of fine-grained category level expert descriptions using LLMs, 2) Few shot category-level PEFT of CLIP and 3) Evaluation on a series of challenging domain shift scenarios.

gration, combining specialized disease identification models with a domain-fine-tuned LLM (ChatGLM3-6B) to provide tailored prescriptions using LoRA. This method involved generating $\sim 2,500$ question-answer pairs using 13 crop-related books and querying Llama3.1-405b instruct. AgriClip [10] utilizes contrastive fine-tuning methodology centered on the curation of extensive image-text paired datasets. They leverage LLMs to generate descriptions for each queried image and subsequently train a CLIP-like architecture using a contrastive learning objective for image and disease classification.

Domain-specific models often face significant operational challenges in terms of requiring high computational resources for the training of large number of parameters of the base LLMs or specialized classification models separately. Domain specific VLM methods [3, 13, 18, 21] rely on the generation of individual image captions or large-scale pre-training question-answer datasets significantly affecting the scalability of data generation in real world settings. This increases the resource demands and cost of generating image descriptions, scaling directly with the number of images queried. These limitations highlight a compelling need for a more computationally efficient and domain-adaptive framework that can enhance robustness and generalization without incurring excessive resource demands. As such, the proposed framework directly addresses these challenges by leveraging a PEFT approach and generating rich, category-level text descriptions (instead of image-level captions), thereby improving efficiency and accessibility for specialized agricultural tasks.

3. Method

In this section, we discuss the proposed CLIP LoRA + A framework designed to bridge the domain gap in agricultural tasks under zero and few shot settings. The detailed flow of CLIP LoRA + A framework is shown in Fig. 1. Our proposed framework encompasses three key stages: (i) class specific textual description generation: utilizing LLMs to generate detailed, class-specific textual descriptions by employing structured and task specific prompting methods (as detailed in Section 3.2), (ii) model fine-tuning: adapting the pre-trained CLIP model using PEFT along with the class-specific generated descriptions through our novel fine-tuning approach (described in Section 3.3), and (iii) evaluation: assessing the performance of the adapted models on few-shot classification across real world datasets and two different agricultural tasks (detailed in Section 3.4).

3.1. Problem statement

Consider a labeled dataset $\mathcal{D} = \{(\mathbf{X}_K, C_K)\}_{K=1}^N$, where $\mathbf{X}_K \in \mathcal{X}$ represents the input images, $C_K \in \mathcal{Y}$ are the corresponding labels for all $K \in \{1, \dots, N\}$, and N is the total number of image-label pairs. We generate an exhaustive list of descriptive textual descriptions, Y^C for each class category $C \in \mathcal{Y}$, by prompting a LLM to identify unique visual and phenological characteristics specific to that category as detailed in Section 3.2. A VLM such as CLIP [13], is composed of an image encoder Θ and a text encoder Φ . The model is trained such that the embedding of the image is close to the embedding of its corresponding text, i.e., $\Theta(\mathbf{x}) \approx \Phi(\mathbf{y})$ for an image \mathbf{X} with label \mathbf{C} . Our objective is to enhance the few-shot generalization capabilities of

Task	Prompt Template	Key Emphasis (Visual Cues)
Disease classification	<p>What are the distinguishable characteristics that can be used to differentiate <disease name>disease from other type of <crop name>plant diseases based on just a photo? Produce an exhaustive list of all attributes that can be used to identify the <disease name>disease uniquely. Give emphasis on location of occurrence of disease, phenological stage at which it generally occurs, physical symptoms as seen on plant/leaf. Texts should be of the form "\ <disease name>disease with <characteristic feature>\". Ensure to structure your response as a list of single sentences.</p> <p>What are the distinguishable characteristics that can be used to differentiate <crop name>plant from other types of crops based on just a photo? Produce an exhaustive list of all attributes that can be used to identify the <crop name>plant uniquely. Give emphasis on colour, appearance, shape, texture of leaves and fruits. Texts should be of the form "<crop name>plant with <characteristic feature>". Ensure to structure your response as a list of single sentences.</p>	<p>Location, Phenological stage, Physical symptoms</p>
Crop classification	<p>What are the distinguishable characteristics that can be used to differentiate <crop name>plant from other types of crops based on just a photo? Produce an exhaustive list of all attributes that can be used to identify the <crop name>plant uniquely. Give emphasis on colour, appearance, shape, texture of leaves and fruits. Texts should be of the form "<crop name>plant with <characteristic feature>". Ensure to structure your response as a list of single sentences.</p>	<p>Shape, Texture, Color, Unique appearance</p>

Table 1. LLM prompt templates for task-specific expert textual description generation

CLIP within fine-grained agricultural tasks by fine-tuning both the image and text encoders in a resource efficient way. We structure this adaptation process by first pre-training the model on large open source laboratory dataset such as PVD, and then adapting the VLM on niche real world dataset.

3.2. Class specific expert textual description generation

To create highly discriminative text features for our vision tasks, we generated class-specific textual descriptions for every class category in each dataset, focusing on unique visual characteristics that facilitate inter-class distinction within the task. The design of the prompt used for text generation was specifically tailored to the respective tasks. For plant disease classification task, the prompt emphasized domain-specific diagnostic attributes, including the location of disease occurrence, the phenological stage at which symptoms typically manifest, and the observable physical symptoms on the plant or leaf. For crop classification task, the generated text centered on identifying features such as the shape, texture, color and unique visual appearance of the leaves and fruits. Table 1 shows respective prompt used to query the LLM. Here, <disease name> and <crop name> is the class name for the classes in the plant disease dataset and crop classification dataset respectively.

The LLM produces n_c descriptions for each class category C . This results in a set of descriptions Y^C for each category. We manually assessed the accuracy of the generated textual descriptions for all classes within the PVD, Pdoc, and the nine classes of the fieldPlant dataset using online sources. This fact-checking process confirmed the correctness of approximately 97% of the generated texts. However, acknowledging that this manual vetting approach is not scalable to larger datasets, we rely on the empirical performance metrics to validate the overall utility of the generated text corpus.

3.3. VLM finetuning

The architecture of original CLIP model leverages a training paradigm involving paired image and caption data. For our specific scenario, however, we utilize a collection of images (\mathbf{X}) and a corresponding set of class-specific textual descriptions (\mathbf{Y}^C) for each class C in our training data. We resolve this data format difference by stochastically pairing each image with a randomly chosen text sample from its respective class during training. To ensure compatibility with the input token constraints of the open CLIP architecture, we strategically sample from the available descriptions to maintain a sequence length within the 77-token limit. Crucially, we cannot directly implement the standard batch-level cross-entropy loss utilized by CLIP, which typically treats the image-text pair as the sole positive example and all other batch texts as negatives. This constraint arises because a single batch in our setup may contain multiple image-text pairs belonging to the identical category. The modification to the objective function necessary to account for this is detailed below.

For each fine-tuning step, we construct a batch of B elements, where each element is an image-text pair $(\mathbf{x}_i, \mathbf{y}_i)$ for $i = 1$ to B . The similarity score resulting from passing image \mathbf{x}_i and text \mathbf{y}_j through the CLIP encoders is denoted by \mathcal{M}_{ij} . $c(i)$ is assigned as the class label for the i -th pair. To identify positive samples correctly, \mathcal{S}_i is defined as the set of indices j such that the category of the j -th pair matches the category of the i -th pair; mathematically, $\mathcal{S}_i = \{j \mid c(j) = c(i)\}$. The resulting loss function for the image component is subsequently expressed as:

$$\mathcal{L}_{\text{image}} = -\frac{1}{B} \sum_{i=1}^B \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \log \frac{\exp(\mathcal{M}_{i,j})}{\sum_{r=1}^B \exp(\mathcal{M}_{i,r})} \quad (1)$$

$$\mathcal{L}_{\text{text}} = -\frac{1}{B} \sum_{j=1}^B \frac{1}{|\mathcal{S}_j|} \sum_{i \in \mathcal{S}_j} \log \frac{\exp(\mathcal{M}_{i,j})}{\sum_{r=1}^B \exp(\mathcal{M}_{i,r})} \quad (2)$$

The overall loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{image}} + \mathcal{L}_{\text{text}} \quad (3)$$

The goal of the objective function is to accumulate the cross-modal similarity for all image and text pairs that share the identical class label within the sampled batch.

We fine-tuned CLIP using PEFT technique specifically Low-Rank Adaptation (LoRA) [6] that significantly reduces the number of trainable parameters required to adapt large pre-trained models to specific downstream tasks. The core hypothesis of LoRA is that the intrinsic rank of the incremental weight update ($\Delta \mathbf{W}$) needed for task adaptation is substantially lower than the rank of the original full weight matrix (\mathbf{W}). The adaptation matrix $\Delta \mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ is modeled as the product of two low-rank matrices, $\mathbf{B} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times d_2}$, where $r \ll \min(d_1, d_2)$ and γ is a scaling factor. The forward pass incorporating the applied LoRA module for an input \mathbf{x} , a hidden state \mathbf{h} is as follows:

$$\mathbf{h} = \mathbf{W}\mathbf{x} + \mathbf{B}\mathbf{A}\mathbf{x} \quad (4)$$

During the fine-tuning process, only the low-rank matrices \mathbf{A} and \mathbf{B} are subject to training and gradient updates, while the parameters of the original pre-trained model (\mathbf{W}) remain frozen. For inference, the adapted matrices are integrated directly with the original weights as shown in equation 4, allowing the model to produce the final result without introducing any additional latency compared to the standard model. Similar to the original LoRA paper [6], we applied the low-rank matrices on the query, key and value matrices of all the attention module of CLIP architecture. The proposed framework is illustrated in Algorithm 1.

3.4. Inference

To overcome the limitations of generic template: ‘‘a photo of <class>’’ for each class k and to evaluate the fine-tuned model on unseen or seen classes in zero and few shot settings respectively, our primary inference strategy utilizes the exhaustive set of LLM-generated textual descriptions $Y^k = \{y_1^k, y_2^k, \dots, y_{l_k}^k\}$ associated with each category k . For any given input image \mathbf{x} , we denote the cosine similarity between the image embedding and the m -th text description of class k as \mathcal{S}_m^k . The final classification decision \hat{k} is determined by aggregating these individual scores using a softmax-normalized probability distribution. The predicted class is the one that maximizes the average probability across its corresponding textual attributes. By calculating the average similarity between the image and the diverse texts corresponding to each class, the model effectively leverages a broader semantic context. The proposed inference is illustrated in Algorithm 2.

Algorithm 1 : Parameter efficient VLM adaptation via expert description

Require: Pre-trained CLIP model (encoders Θ, Φ), Large Laboratory Dataset \mathcal{D}_{lab} (e.g., PVD), Real-world Dataset $\mathcal{D}_{\text{real}}$, Class Categories \mathcal{Y} .

Ensure: Adapted agricultural CLIP model.

- 1: **Class-Specific Textual Description Generation:**
- 2: **for** each class category $C \in \mathcal{Y}$ **do**
- 3: Define Task T
- 4: Formulate task-specific prompt P_T based on Table 1
- 5: Query LLM: $Y^C \leftarrow \text{LLM}(P_T, C)$
- 6: **end for**
- 7: **VLM Fine-Tuning:**
- 8: Initialize LoRA layers A, B for $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ in Θ, Φ
- 9: Pre-train Θ, Φ on \mathcal{D}_{lab}
- 10: **while** not converged on $\mathcal{D}_{\text{real}}$ **do**
- 11: Sample batch of size B from $\mathcal{D}_{\text{real}} : \{(\mathbf{x}_i, c(i))\}_{i=1}^B$
- 12: **for** each $i \in \{1, \dots, B\}$ **do**
- 13: Select random $y_i \in Y^{c(i)}$ {image-text pairing}
- 14: **end for**
- 15: Compute similarity matrix $\mathcal{M}_{ij} = \Theta(\mathbf{x}_i) \cdot \Phi(\mathbf{y}_j)$
- 16: Identify positive indices: $\mathcal{S}_i = \{j \mid c(j) = c(i)\}$
- 17: Calculate class-aware contrastive loss \mathcal{L} using Eq. 3
- 18: Update LoRA parameters A, B via $\nabla_{A,B}(\mathcal{L})$
- 19: **end while**
- 20: **return** Fine-tuned model (Θ, Φ)

Algorithm 2 : Inference

Require: Input image \mathbf{x} , set of test classes $\mathcal{K}_{\text{test}}$, LLM-generated text sets $\{Y^k\}_{k \in \mathcal{K}_{\text{test}}}$

Ensure: Predicted class \hat{k} .

- 1: // Extract visual embedding
- 2: $\mathbf{z}_v \leftarrow \Theta(\mathbf{x})$ {Image Encoder}
- 3: **for** each class $k \in \mathcal{K}_{\text{test}}$ **do**
- 4: $l_k \leftarrow |Y^k|$ {Number of descriptions for class k }
- 5: **for** each text description $y_m^k \in Y^k$ **do**
- 6: $\mathbf{z}_{t,m}^k \leftarrow \Phi(y_m^k)$ {Text Encoder}
- 7: $\mathcal{S}_m^k \leftarrow \text{cosine_similarity}(\mathbf{z}_v, \mathbf{z}_{t,m}^k)$
- 8: **end for**
- 9: **end for**
- 10: // Compute softmax-normalized probabilities per class
- 11: **for** each class $k \in \mathcal{K}_{\text{test}}$ **do**
- 12: $P_k \leftarrow \frac{1}{l_k} \sum_{m=1}^{l_k} \frac{\exp \mathcal{S}_m^k}{\sum_{p \in \mathcal{K}_{\text{test}}} \sum_{q=1}^{l_p} \exp \mathcal{S}_q^p}$
- 13: **end for**
- 14: // Selection of predicted category
- 15: $\hat{k} \leftarrow \arg\max_k (P_k)$
- 16: **return** \hat{k}

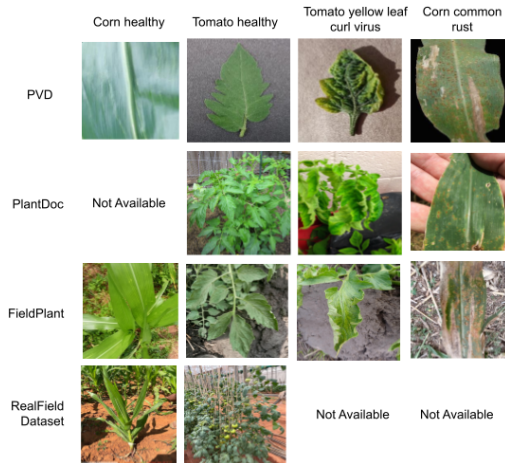


Figure 2. Example of images from PVD, PlantDoc, FieldPlant and RealField datasets showing variations in data collection environment for the same category.

4. Experiments

4.1. Datasets

In our study, we utilized three open source plant image datasets i.e., the PlantDoc (Pdoc) [14], the Plant Village Dataset (PVD) [8], and FieldPlant [9] and one self collected real world crop classification dataset named RealField dataset. The open source datasets are structured for plant disease detection task. These datasets collectively provide a comprehensive benchmark across different environmental conditions, geographic variation and limited labels per class. The PVD is a collection of high-resolution images capturing a single, isolated leaf taken in a controlled, laboratory environment, leading to clean visual separation of classes. Conversely, RealField, Pdoc and FieldPlant feature complex, real-world field images, often showing the entire plant, including multiple leaves, stems and fruits. RealField, Pdoc and FieldPlant datasets showed significant domain variability due to inconsistent lighting, background clutter and occlusion, presenting a more challenging classification scenario. Notably, the Pdoc and FieldPlant datasets are characterized by a limited number of labels per class for disease classification task, which reflects a common constraint in real-world agricultural data acquisition. The RealField dataset contains two types of crops, i.e., 130 images of tomato crop grown in poly house setup and 62 images of maize crop grown in agricultural field in Bengaluru, Karnataka, India. These two crop images were captured using android mobile phone with 32 megapixel camera over different phenological stages at an interval of 15 days starting from 30 days after planting till harvesting, thus, introducing variations with respect to different growth stages, lightning conditions and other environmental conditions during the

crop life cycle. Table 2 summarizes the individual configuration, totaling 62249 images across all the four datasets.

To test the framework for crop classification task, we derived a secondary set of datasets from PVD, Pdoc and FieldPlant, focused purely on crop classification. This transformation was achieved by aggregating all diseased and healthy images belonging to a particular crop type into a single unified class. For example, all images labeled “Apple Scab”, “Apple Black Rot”, and “Apple Healthy” were combined under the single class “Apple”. Refer to Table 2 for details on the number of classes per dataset.

4.2. Implementation details

For synthesizing category-level textual descriptions used in training, we utilized the Gemini 2.5 Pro API with temperature parameter set to 0 to get consistent and reproducible text descriptions across runs. We adapted the pre-trained CLIP ViT-B/16 model using PEFT [23]. Training was conducted for 100 epochs, incorporating early stopping, which was configured with a patience of 5 epochs and a minimum change ($\min \Delta$) in validation accuracy set to 0. The optimal learning rate was empirically determined via validation and set to 2×10^{-3} across all fine-tuning experiments.

The CLIP architecture comprises of an image encoder and a text encoder, both based on the Transformer architecture, each concluding with a linear projection layer. While a standard approach is to apply LoRA to all weight matrices in both the encoders, we selectively applied it to conserve computational resources and prevent overfitting in the few-shot setting. LoRA modules were employed by applying the low-rank matrices only to the query (**Q**), key (**K**) and value (**V**) matrices within the self-attention mechanisms of both the vision and text encoders. The rank was set conservatively at $r = 2$. To further regularize the input to the LoRA modules, a dropout layer with a probability of $p = 0.25$ was incorporated. The hyper-parameters were fixed across all fine-tuning experiments for consistency.

4.3. Baselines

In this section, we discuss the various methods for which we compare classification accuracy in both zero and few-shot settings.

CLIP [13] refers to pre-trained CLIP tested with “a photo of a [class]” texts like the original paper.

CLIP LoRA [23] refers to finetuning CLIP using PEFT tested with “a photo of a [class]” texts like the original paper.

CLIP LoRA + A is our proposed method where we fine-tune CLIP using PEFT and category level cotrastive loss utilizing category level expert text descriptions generated by LLM.

Dataset	Total images	Task I: Disease classification classes	Task II: Crop classification classes	Environment	Data challenge
Plant Village Dataset (PVD)	54,303	38	14	Controlled	Large-scale, Clean
PlantDoc (Pdoc)	2,598	28	13	Real world field	Limited Samples, High Variability
FieldPlant	5,156	26	3	Real world field	Limited Samples, High Variability
RealField	192	-	2	Real world poly house/field	Limited Samples, High Variability
Total	62249				

Table 2. Summary of datasets used. Detailing the original disease classification task (Task I) and the derived crop classification task (Task II) configurations.

Shots	Task	Disease classification		Crop classification	
	Method	Pdoc	FieldPlant	Pdoc	FieldPlant
0	CLIP LoRA	40.53	40.89	73.75	89.65
	CLIP LoRA + A (ours)	40.71	40.79	73.21	88.65
4	CLIP LoRA	61.64	43.94	80.83	97.22
	CLIP LoRA + A (ours)	63.12	81.55	82.2	98.74
8	CLIP LoRA	62.54	50.25	81.67	99.24
	CLIP LoRA + A (ours)	65.53	86.88	83.05	99.74
full	CLIP LoRA	70.04	55.64	83.75	99.78
	CLIP LoRA + A (ours)	86.14	88.46	88.63	99.77

Table 3. Classification accuracy (%) on the challenging real-world datasets (Pdoc and FieldPlant) for both disease and crop classification tasks. All models were initialized after being fine-tuned on the large, controlled PVD laboratory dataset.

Shots	Task	Disease classification			Crop classification			
		Method	PVD	Pdoc	FieldPlant	PVD	Pdoc	FieldPlant
0	Clip		10.18	27.97	20.08	14.78	25.42	81.53
	CLIP LoRA		10.71	37.25	22.98	14.95	39.71	79.29
	CLIP LoRA + A (ours)		10.71	37.25	22.98	14.95	39.71	79.29
full	Clip		10.41	30.57	35.72	38.36	29.24	87.25
	CLIP LoRA		98.13	76.11	82.58	99.96	72.08	99.45
	CLIP LoRA + A (ours)		98.15	77.73	87.99	99.12	75.91	99.75

Table 4. Classification accuracy (%) under zero shot (0) and utilizing full training data (full) for finetuning without any pretraining using laboratory dataset.

5. Results

In this section, we compare our method, CLIP LoRA + A (where ‘A’ signifies augmentation using LLM-generated expert text descriptions), against the baselines, standard CLIP and CLIP LoRA, and evaluate its performance under various specialized settings. We demonstrate the superiority of our framework across different agricultural vision tasks and domain shifts. We further show that our method maintains performance advantages even under difficult evaluation scenarios, including performance without pre-training on the large laboratory PVD dataset, zero-shot performance on a real-world novel field dataset, and improved zero-shot and few-shot performance with pre-training on the PVD dataset.

5.1. Comparison with baselines

Table 3 evaluates the performance of the models under zero-shot (0-shot), few-shot (4 and 8-shot) and full-training data (full) settings on the challenging real-world datasets, Pdoc and FieldPlant. These models were initially pre-trained on the large, controlled PVD laboratory dataset using LoRA-based PEFT. The results demonstrated that combining PVD pre-training with our expert textual descriptions significantly boosts performance in the few-shot regime, while zero-shot accuracies remained nearly identical between the two methods, as the influence of expert descriptions primarily emerged after target-domain fine-tuning. At 4 and 8 shots, CLIP LoRA + A consistently surpassed the CLIP LoRA baseline across all the four datasets, demonstrating exceptional transferability with minimal field-specific labels. Furthermore, as shown in Table 3, CLIP LoRA + A achieved the highest overall classification accuracy when fine-tuned on the full training dataset.

5.2. Effect of pretraining using laboratory dataset

Table 4 presents the comparative results for zero-shot (0) and utilizing full training data (full) fine-tuning without any intermediate agricultural pre-training (i.e., the models begin from their original LAION pre-training). In the zero-shot setting, both the CLIP-LoRA baseline and our proposed CLIP LoRA + A achieved identical results across all six tasks. This is expected, as without any fine-tuning data, the performance relies entirely on the quality of the model’s general pre-trained knowledge. When all available training data is used i.e full finetuning, there is a significant improvement in accuracy as compared to the zero shot. This suggests that the LLM-generated expert text descriptions (A) provides crucial, fine-grained semantic supervision, particularly valuable for complex, noisy field images. However, a comparison with the results in Table 3 revealed a markedly superior overall performance when the models are fine-tuned on top of an in-domain pre-trained source, such as the large, controlled PVD laboratory dataset. This highlights the foundational role of intermediate domain pre-training in agricultural vision, which significantly boosts subsequent performance on limited, real-world field datasets.

5.3. Evaluation on novel field

Table 5 evaluates the zero-shot performance of models pre-trained on various source datasets when tested on the RealField dataset, simulating deployment in a completely novel geographical domain. This setup directly assessed the domain generalization capability. Zero shot accuracy of the baseline CLIP model pre-trained on LAION 400M achieved 79.83% similar to the proposed CLIP LoRA + A model (78.06%) due to the same reason mentioned in Section 5.1. When the proposed model was pre-trained on domain-specific data, the zero shot performance improved substantially by an average increment of 17.53%. The highest zero-shot accuracy on the novel RealField dataset was achieved by CLIP LoRA + A, when pre-trained on Pdoc (99.48%) and FieldPlant (98.77%) since these two datasets consist of real world field crops unlike PVD dataset that consists of single leaf laboratory dataset. Crucially, CLIP LoRA + A consistently outperformed CLIP LoRA and standard CLIP when pre-trained on the agricultural datasets, demonstrating that the rich textual context gained from our framework facilitated superior generalization to entirely unseen target domains.

5.4. Evaluation on novel classes/crops

The fieldPlant disease and crop classification dataset consisted of 16 and 1 novel classes respectively, as compared to PVD dataset. The results discussed in Table 3 implicitly address performance on novel classes, as the pre-training on PVD involved a different set of categories and a different domain (controlled lab versus real field) compared to the target FieldPlant dataset. The strong performance of CLIP LoRA + A in the few-shot settings (4 and 8 shots) for both disease and crop classification tasks on the fieldPlant dataset confirms the framework’s ability to effectively transfer knowledge and rapidly adapt to novel, fine-grained categories with minimal labeled examples.

5.5. Resource requirements

Our framework is highly resource-efficient, utilizing PEFT via LoRA in finetuning CLIP architecture. The low-rank update matrices signifies that only a small fraction of the model parameters are trained, significantly reducing memory and computational overhead compared to full network fine-tuning. In our case, out of 150 million trainable parameters of openClip Vit B/16 architecture, we trained only 0.12% of the parameters. Additionally, generating category-level expert text descriptions by using LLM prompting further enhances efficiency in both time and cost by avoiding the need to query the LLM for individual image captions for every image in the dataset, which is a common and costly bottleneck in other agricultural VLM methods. The cost of using the Gemini-2.5-pro API to query text descriptions for a dataset with 20-40 classes (such as PVD) via

Method	Source Domain			
	LAION 400M	PVD	Pdoc	FieldPlant
CLIP	79.83	85.94	95.98	96.98
CLIP-LoRA	79.06	86.68	98.64	98.44
CLIP LoRA + A (ours)	78.06	88.54	99.48	98.77

Table 5. Zero-shot classification accuracy (%) on the RealField dataset after the models have been fine-tuned on various source domains (LAION 400M, PVD, Pdoc and FieldPlant)

our method is about \$1-\$5. All training was feasible on a single 48 GB RTX 8000 Nvidia GPU, confirming the practical accessibility of the method for real-world agricultural deployment.

6. Conclusion

We present a method to improve the zero-shot and few-shot classification performance of VLMs using expert category-level textual description generated by LLMs on fine-grained agricultural domain tasks. CLIP LoRA + A eliminates the need for costly image-caption datasets. It integrates expert phenological and morphological descriptors with a novel category-level contrastive loss and Parameter-Efficient Fine-Tuning (PEFT) via LoRA, ensuring effective adaptation with minimal computational overhead. The experimental results demonstrated that our framework consistently outperformed vanilla CLIP and standard CLIP-LoRA baselines. Additionally, our method achieved 99.48% zero-shot accuracy on the novel RealField dataset, proving superior robustness across unseen geographic domains. The findings suggests that category level expert text description priors are equally effective and complementary to visual information for zero shot and few shot classification in agricultural domain.

Acknowledgements

We are grateful to Regional Remote Sensing Centre (RRSC-South), National Remote Sensing Center (NRSC), Indian Space Research Organization (ISRO), Bengaluru for data and research grant. IIIT Bangalore is acknowledged for the infrastructure support. We also acknowledge Mphasis Cognitive Computing Centre of Excellence for the financial assistance.

References

- [1] Deeksha Aggarwal, Sai Shruti Prakhya, and Uttam Kumar. Agriculture crop monitoring for yield estimation with zero-shot fruit detection: A deep learning approach. In *Remote Sensing of Land Cover and Land Use Changes in South and Southeast Asia, Volume 1*, pages 115–132. CRC Press. 1

- [2] Deeksha Aggarwal, Yash Mittal, and Uttam Kumar. Advancing image classification through parameter-efficient fine-tuning: A study on lora with plant disease detection datasets. In *The Second Tiny Papers Track at ICLR 2024*, 2024. 1
- [3] Muhammad Awais, Ali Husain Salem Abdulla Alharthi, Amandeep Kumar, Hisham Cholakkal, and Rao Muhammad Anwer. Agrogpt: Efficient agricultural vision-language model with expert tuning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5687–5696. IEEE, 2025. 2, 3
- [4] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT Press, 2016. 1
- [5] Mohammadreza Haghighat, Alzayat Saleh, and Mostafa Rahimi Azghadi. Multimodal language models in agriculture: A tutorial and survey. *Information Fusion*, page 104042, 2025. 1
- [6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 5
- [7] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2
- [8] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016. 6
- [9] Emmanuel Moupojou, Appolinaire Tagne, Florent Retraint, Anicet Tadonkemwa, Dongmo Wilfried, Hyppolite Tapamo, and Marcellin Nkenlifack. Fieldplant: A dataset of field plant images for plant disease detection and classification with deep learning. *IEEE Access*, 11:35398–35410, 2023. 6
- [10] Umair Nawaz, Awais Muhammad, Hanan Gani, Muzammal Naseer, Fahad Shahbaz Khan, Salman Khan, and Rao Anwer. Agrclip: Adapting clip for agriculture and livestock via domain-specialized cross-model alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9630–9639, 2025. 3
- [11] Umair Nawaz, Awais Muhammad, Hanan Gani, Muzammal Naseer, Fahad Shahbaz Khan, Salman Khan, and Rao Anwer. Agrclip: Adapting clip for agriculture and livestock via domain-specialized cross-model alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9630–9639, 2025. 1
- [12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2
- [13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2, 3, 6
- [14] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. Plantdoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 249–253. 2020. 6
- [15] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [16] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geolocalization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023. 1
- [17] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, page 3876, 2022. 1
- [18] Changqing Yan, Zeyun Liang, Han Cheng, Shuyang Li, Guangpeng Yang, Zhiwei Li, Ling Yin, Junjie Qu, Jing Wang, Genghong Wu, et al. Cclip-chatglm3: A dual-model approach integrating computer vision and language modeling for crop disease identification and prescription. *Computers and Electronics in Agriculture*, 236:110442, 2025. 2, 3
- [19] Xing Yang, Lei Shu, Jianing Chen, Mohamed Amine Ferrag, Jun Wu, Edmond Nurellari, and Kai Huang. A survey on smart agriculture: Development modes, technologies, and security and privacy challenges. *IEEE/CAA Journal of Automatica Sinica*, 8(2):273–302, 2021. 1
- [20] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023. 2
- [21] Piaofang Yu and Bo Lin. A framework for agricultural intelligent analysis based on a visual language large model. *Applied Sciences*, 14(18):8350, 2024. 2, 3
- [22] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2
- [23] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024. 2, 6
- [24] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022. 2
- [25] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *In-*

ternational Journal of Computer Vision, 130(9):2337–2348, 2022. [2](#)

- [26] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. [2](#)