School of Computer Science

Faculty of Science and Engineering

University of Nottingham

Malaysia

UG FINAL YEAR DISSERTAION REPORT

- Title -

Learning Generalizable Models using CLIP with Seed Level Descriptions

Student's Name: Khalid Al Najjar

Student Number: 20490270

Supervisor's Name: Dr. Iman Yi Liao

Year: 2025

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR OF SCIENCE IN COMPUTER SCIENCE – ARTIFICIAL INTELLIGENCE BSc (HONS) THE UNIVERSITY OF NOTTINGHAM

- Title -

Learning Generalizable Models using CLIP with Seed Level Descriptions

Submitted in May 2025, in partial fulfilment of the conditions of the award of the degrees B.Sc.

- Khalid Al Najjar -

School of Computer Science

Faculty of Science and Engineering

University of Nottingham

Malaysia

I hereby declare that this dissertation is all my work, except as indicated in the text:

Khalid Al Najjar

04/05/2025

Table of Contents

Abstract 3

1 Introduction 4

2 Motivation 5

3 Related Work 5

3.1 AI in agriculture 5

3.2 Interactive ML for Soybean Seed Classification 6

3.3 Machine Vision in Seed Quality Assessment 6

3.4 CLIP Model for Zero-Shot Classification 7

3.5 Transfer Learning for Crop Classification 7

3.6 Context Optimization for CLIP Models 8

4 Description of The Work 9

5 Methodology 10

5.1 Data Handling 10

5.2 Descriptor Generation 10

5.3 Model Preparation 10

5.4 Training Process 10

5.5 Testing 11

5.6 Prompt Evaluation 11

5.7 Heatmap Generation 11

6 Design 12

7 Implementation 13

7.1 Data Handling and Processing 13

7.2 Prompt Generation with OpenAI API 14

7.3 Model Fine-Tuning 14

7.4 Training and Validation 14

7.5 Model Testing 15

7.6 Performance Metrics 15

8 Evaluation 20

8.1 CLIP Zero-Shot Testing 20

8.1.1 Zero-Shot with CLIP Recommended Prompts 20

8.1.2 Zero-Shot with Image Augmentations 21

8.2 CLIP Model Training with ChatGPT Prompts 23

8.2.1 Training Using Simple Class Prompts 23

8.2.2 Training with More Complex Prompts 27

8.2.3 Prompt Refinement Process 30
9 Summary and Reflections 32
References 33

Abstract

Oil palm is a vital crop worldwide, playing a critical role in the agricultural economy of many tropical countries. Its seeds facilitate a wide range of industries, from food production to cosmetics and biofuels. The success of oil palm cultivation begins with ensuring a good quality of seeds. To ensure high yields, seed quality must be carefully managed, beginning with the germination stage. Traditional seed classification methods rely on human expertise which can be subjective, labor-intensive, and time-consuming. Automating this process can enhance efficiency, increase accuracy and reduce costs. However, distinguishing between “good” and “bad” oil palm seeds is challenging due to subtle visual differences such as color, texture and shape. Furthermore, environmental factors like lighting and angles further complicate classification, necessitating a model capable of strong generalization. Incorporating image captioning into the classification pipeline adds an interpretative layer to the model’s outputs. By matching images with a set of descriptive keywords, the model learns to associate textual descriptors with visual seed properties. This approach should not only improve the model’s classification performance but also make its predictions more transparent to the user.

1 Introduction

The classification of germinated seeds has always been a necessary step in improving agricultural productivity, particularly for high-value crops like oil palms [13]. Accurate seed classification enables better resource allocation, higher germination success rates, and improved yield outcomes. Traditionally, this process has relied on manual inspection, which lacks scalability and is vulnerable to human error and subjectivity.

In recent years, artificial intelligence (AI) and deep learning have shown promise in automating complex tasks in agriculture, including plant disease detection, crop monitoring, and seed viability assessment [11]. However, the application of AI to seed classification presents specific challenges. Many deep learning-based models demonstrate high performance on controlled datasets but often struggle to generalize to new conditions or environments due to overfitting or sensitivity to noise, such as lighting or image quality.

To address these limitations, vision-language models (VLMs) like CLIP (Contrastive Language-Image Pretraining) have emerged as a promising direction. CLIP models jointly learn image and text representations, enabling zero-shot or prompt-based classification using descriptive natural language. CLIP, which was trained on 400 million image-text pairs, maps both images and textual descriptors into a shared embedding space, allowing it to capture visual inputs with textual prompts directly. This makes it particularly attractive for tasks where labeled data is limited or where robustness across varying conditions is essential.

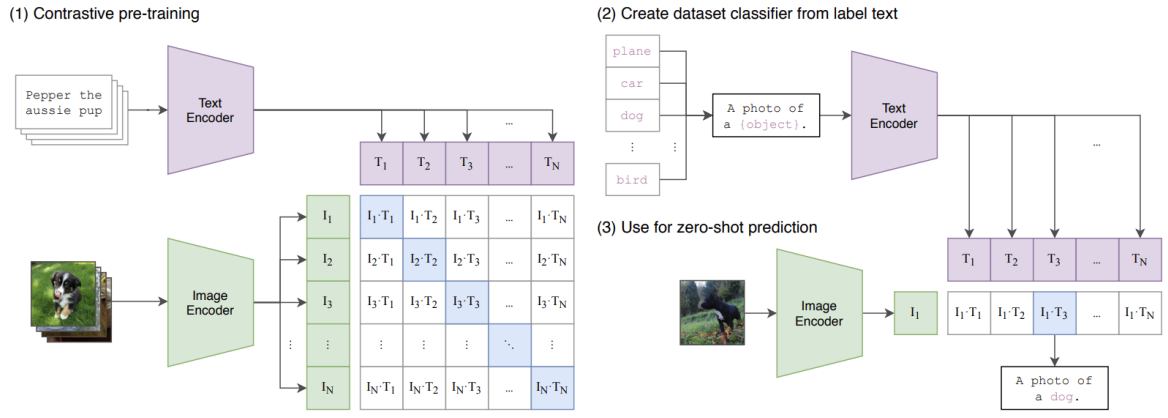

Figure 1. Summary of CLIP architecture

This project investigates whether integrating natural language descriptors generated by large language models (LLMs), such as ChatGPT, can improve classification performance and generalizability in a seed quality assessment task. By combining an automated preprocessing pipeline, LLM-driven descriptor generation, and CLIP, this work aims to produce a robust model for germinated oil palm seed classification.

2 Motivation

The motivation for this project stems from the limitations of manual and traditional automated methods for classifying oil palm seeds. Manual classification is labor-intensive, time-consuming and prone to human error. This traditional approach often suffers from inconsistencies, as the outcomes can vary depending on the individual's skill and experience. Automating the seed classification process can streamline the process, increase accuracy, and reduce costs, addressing critical inefficiencies in the agricultural supply chain.

Visually classifying oil palm seeds into "good" and "bad" is challenging due to subtle visual variations in appearance, such as color texture, and shape [12]. These variations can be difficult to detect, even for a skilled human operator. Moreover, environmental factors such as lighting inconsistencies, camera positioning, and image quality add further complexity to classification tasks, making generalization across datasets a key challenge.

This project seeks to overcome these limitations by leveraging recent advances in multimodal AI. Specifically, by coupling image features with contextually rich textual descriptions. The hypothesis is that incorporating natural language guidance using LLMs can help vision-language models better distinguish between seed classes and perform robustly in varied conditions, making the system more deployable in practical agricultural environments.

3 Related Work

3.1 AI in agriculture

AI has made significant advances in agriculture as highlighted in the work of Luís Santos [1], the author explores the diverse applications of deep learning (DL) in addressing agricultural challenges including crop disease detection, yield estimation, soil quality assessment and precision farming. Techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been employed to analyze complex imagery and data, leading to improved decision-making and resource management. The review shows the potential of these models to revolutionize agricultural practices by enhancing efficiency and reducing costs. This work highlights the broad utility of AI solutions in agriculture, providing valuable context for the development of systems like our oil palm seed classification project.

3.2 Interactive ML for Soybean Seed Classification

de Medeiros et al. [2] introduced an approach that combines interactive and traditional machine learning methods to classify soybean seeds and seedlings based on their appearance and physiological potential. By utilizing computer vision and AI algorithms, the researchers aimed to reduce subjectivity and enhance the efficiency of the analysis process. They developed models using low-cost models and free-access software, achieving high performance with overall accuracy ranging between 0.92 and 0.99 for the interactive machine learning model, and greater than 0.90 overall accuracy in the independent validation set for external classification. The findings indicate a strong correlation between the appearance of soybean seeds and their physiological performance, suggesting that the proposed method can rapidly and objectively classify soybean seed vigor.

3.3 Machine Vision in Seed Quality Assessment

Research by Tu Ke-ling et al. [3] aimed to enhance the selection process of pepper seeds by distinguishing high quality seeds from low quality ones using machine vision and classification algorithms. Recognizing that seed Vigor correlates significantly with physical attributes like color and size, the study identified several physical features as potential predictors of seed quality. The research employed image recognition software to automate the assessment of these features in 400 images of pepper kernels. They applied binary logistic regression and a multilayer perceptron (MLP) neural network to develop predictive models for seed germination. The study found that certain color features (R, a*, brightness), dimensions (width, length, projected area) and weight positively correlated with seedling fresh weight. The MLP neural network demonstrated high predictive accuracy, with a germination percentage of 79.1% and a selection rate of 90.0% in the calibration set. These findings suggest that machine vision combined with classifier models can effectively predict seed germination based on physical characteristics, offering a cost-effective and labor-saving approach for quality control in seed selection. The identification of specific physical features (e.g. color metrics, size dimensions) as predictors of seed quality in pepper seeds may inspire similar feature selection criteria for our oil palm seed dataset, potentially enhancing the accuracy of our classification model.

3.4 CLIP Model for Zero-Shot Classification

Building on these agricultural applications, our work also draws from foundational research in computer vision and multimodal learning. The CLIP model that forms the basis of our approach was introduced by Radford et al. [5], who demonstrated how contrastive learning between images and text can create powerful visual representations without class-specific labels.

Unlike traditional CNN-based classifiers that require extensive labeled examples of each class, CLIP can generalize to new concepts using only textual descriptions. This zero-shot capability is particularly valuable for our limited dataset scenario, where obtaining large quantities of labeled oil palm seeds is challenging. Additionally, CLIP's multimodal nature allows us to leverage domain expertise encoded in natural language descriptors, potentially capturing subtle characteristics that distinguish healthy seeds from unhealthy ones. The model's architecture, consisting of separate image and text encoders that project into a shared embedding space, provides an ideal foundation for our approach of using detailed seed-quality descriptors to guide classification decisions.

3.5 Transfer Learning for Crop Classification

Suh et al. [6] explored transfer learning for crop classification under field conditions, showing that pretrained models can generalize well to new agricultural datasets. Their research focused on distinguishing sugar beet plants from volunteer potato plants in actual field settings, where varying lighting conditions, occlusions, and growth stages present significant challenges.

Particularly relevant to our work, Suh et al. demonstrated that transfer learning methods were robust against environmental variations, maintaining consistent performance across different field conditions. Their findings justify our use of CLIP, a pretrained vision-language model, to adapt to our oil palm seed classification task. By leveraging CLIP's pretrained weights and adapting them to our specific domain, we can benefit from the rich visual representations the model has already learned while tuning it to recognize the specific characteristics of oil palm seeds. The authors' success in handling variable field conditions also suggests that our approach may be effective in addressing the lighting and imaging variations present in seed classification scenarios.

3.6 Context Optimization for CLIP Models

Zhou et al. [7] proposed Context Optimization (CoOp) an approach to adapt CLIP-like visual models for downstream image recognition through learnable context. Traditional prompt engineering for CLIP models often relies on hand-crafted prompts like “a photo of a [CLASS]”, where the class token is replaced by the actual class name like “car” or “plane”, which may not fully leverage the model’s capabilities.

The authors evaluated CoOp across 11 datasets spanning various recognition tasks, their findings showed that CoOp was able to beat hand-crafted prompts by a significant margin (e.g., improving accuracy by 16.5% on average in the few-shot scenario), demonstrating the importance of context in guiding visual-language models. This performance gain was achieved with minimal additional parameters and training data.

For our oil palm seed classification task, CoOp’s findings suggest that optimizing the textual context surrounding class descriptors could substantially improve performance. The success of CoOp validates our hypothesis that enhancing textual descriptions can improve CLIP's generalizability and performance on specialized classification tasks like ours.

4 Description of The Work

This project focuses on developing a model that can distinguish between “good” and “bad” germinated oil palm seeds using CLIP. The system combines both visual and textual information to improve its classification accuracy and generalizability.

The main objectives of this task include generating meaningful descriptors for both seed classes (good/bad) using ChatGPT, eliminating manual descriptor creation while ensuring consistency and relevance in the training process.

These descriptors will be used in a CLIP-based pipeline that simultaneously processes both visual features from seed images and textual descriptors, allowing the model to learn multimodal representations of each class.

The model will then be fine-tuned on a primary dataset and its performance will be evaluated on multiple test datasets captured under varying lighting conditions, to assess generalizability and robustness to environmental variations [14].

Score-CAM (Score-weighted Class Activation Mapping) [4] will be implemented into the pipeline to generate visual heatmaps highlighting regions of highest attention in the images, providing interpretability by revealing which visual features most strongly influence classification decisions [15] [16].

This approach leverages the strength of modern vision-language models to create a practical solution for oil palm seed quality classification that can adapt to different imaging conditions, as well as provide interpretable insights into its classification logic.

5 Methodology

5.1 Data Handling

The process begins with dataset preparation, where images are organized into two classes: “good” and “bad” germinated oil palm seeds. This dataset is then split into a training set (80%) and a validation set (20%). Data augmentation techniques, such as rotation and color jitter, are applied to the training set to improve generalization. This augmentation strategy is supported by research from Perez and Wang [8], which demonstrated that such techniques improve model generalizability across environmental conditions, particularly crucial for our cross-dataset evaluation under different lighting conditions.

5.2 Descriptor Generation

We leverage ChatGPT's language capabilities to automatically generate a predefined number of detailed textual descriptors for both seed classifications. This automated approach eliminates subjective bias in descriptor creation while ensuring coverage of distinguishing seed characteristics. The resulting descriptors are then structured into lists to be used for CLIP’s text encoder.

5.3 Model Preparation

The system employs a pre-trained CLIP model as its foundation, utilizing its dual-encoder architecture to process visual and textual information in parallel. To preserve CLIP’s foundational knowledge, while adapting it to our specific classification task, we implement a progressive layer unfreezing strategy during fine-tuning. This approach aligns with Howard and Ruder’s [9] findings, which show that selectively unfreezing layers, rather than fine-tuning the entire model at once, helps prevent catastrophic forgetting and enables effective domain adaptation.

5.4 Training Process

During training, each image is compared with all text prompts from both seed categories using CLIP’s similarity scoring mechanism. The similarity scores are averaged for each class, and the class with the higher average score is selected as the prediction. This process is repeated for a defined number of epochs.

During training, key metrics such as Cross-entropy loss, precision, accuracy, and F1 score are tracked per epoch, these metrics are stored for visualization and analysis.

Validation is done after each epoch, where the saved model is run on the unseen validation set, to ensure good generalization, and avoid overfitting the model on the training set. The model with the lowest validation loss is saved as the best-performing version.

The model will be fine-tuned by using the unfrozen transformer layers along with the input and output processing layers. The learning process is managed using an optimizer with a learning rate scheduler to ensure stable convergence.

5.5 Testing

After training, the model is evaluated on three test sets: the primary test set, the NormalRoomLight set, and the Lightbox set. These evaluations assess the model's generalizability to unseen data and varying lighting conditions. For each set, a classification report and a confusion matrix are generated to analyze the model’s performance, highlight misclassifications, and identify potential weaknesses in distinguishing between the two seed types.

5.6 Prompt Evaluation

We evaluate the effectiveness of each prompt by analyzing its individual contribution to the classification outcomes. This enables us to identify the most informative prompts, and the potentially less effective ones, which improves our overall model performance.

5.7 Heatmap Generation

To enhance model interpretability, we generate heatmaps that highlight the most influential regions in the classification decision. These visualizations provide insight into the model’s attention patterns and support qualitative analysis of its decision-making process.

6 Design

The design of this system is centered around overcoming three main challenges: limited labelled data, subtle visual differences between classes, and the need for interpretability in classification decisions.

To address the limited data problem, the system leverages the CLIP model, which is inherently well-suited for few-shot and zero-shot learning scenarios. By utilizing textual descriptions as class representations, CLIP reduces reliance on large volumes of labeled training data. Xiaoyi Dong et al. [10] demonstrated that with appropriate hyperparameter tuning, CLIP achieves state-of-the-art fine-tuning performance on ImageNet, surpassing many supervised models, making it a suitable choice for our task.

To manage the challenge of subtle visual differences between “good” and “bad” seeds, the system uses detailed textual prompts that describe class-specific visual features. These prompts are generated via ChatGPT and are carefully crafted to emphasize class-specific characteristics. The prompt generation is integrated via an API, ensuring consistency, relevance, and proper formatting compatible with the CLIP model’s tokenizer.

Interpretability is essential to underrating the model’s decisions. To address this, the design includes a form of prompt effectiveness analysis, which assesses the contribution of each textual prompt to classification outcomes. Additionally, Score-CAM based heatmaps are generated, which offer a look into visual explanation of the classification process.

This model design not only enables effective classification under constrained data conditions but also offers human-understandable explanations for the model’s predictions.

7 Implementation

7.1 Data Handling and Processing

The training dataset is loaded from an image folder structured into “GoodSeed” and “BadSeed” subfolders. It is randomly split into an 80/20 training/validation split using the ‘random_split’ function from the ‘torch.utils.data’ library, ensuring that each run produces a unique split.

Image augmentations are applied to the training set using ‘torchvision.transforms.v2’, including Resize, RandomCrop, ColorJitter, RandomVerticalFlip, and more. These augmentations should help the model to generalize beyond the limited training set and reduce overfitting.

```
training_transform = v2.Compose([
    v2.Resize((256, 256), interpolation=v2.InterpolationMode.BICUBIC),
    v2.RandomResizedCrop(224, scale=(0.7, 1.0), ratio=(0.75, 1.33)),
    v2.RandomRotation(degrees=30),
    v2.RandomHorizontalFlip(p=0.5),
    v2.RandomVerticalFlip(p=0.3),
    v2.ColorJitter(brightness=0.4, contrast=0.4, saturation=0.4, hue=0.1),
    v2.RandomAffine(degrees=15,translate=(0.1, 0.1),scale=(0.9, 1.1),shear=10),
    v2.RandomErasing(p=0.3, scale=(0.02, 0.2), ratio=(0.3, 3.3)),
    v2.GaussianBlur(kernel_size=3, sigma=(0.1, 2.0)),
```

Figure 2. Image augmentations used for training set

Figure 3. Augmented healthy seed image

Figure 4. Augmented unhealthy seed image

7.2 Prompt Generation with OpenAI API

Descriptive prompts for both seed classes are generated using an API call to ChatGPT. The prompts are tailored to highlight distinguishing features between healthy and unhealthy seeds. Each prompt is tokenized using CLIP's tokenizer and encoded with the text encoder to produce class-specific embeddings, which will be used for similarity comparison during training and testing.

```
bad_seed_prompts = ['Dark, shriveled appearance with noticeable discoloration',
                    'Presence of mold or fungal growth on the seed surface',
                    'Cracked or split seed coat with visible damage']

good_seed_prompts = ['Bright, vibrant color with a glossy appearance',
                     'Firm, plump seed shape that is free of wrinkles',
                     'Intact, smooth testa with no visible blemishes or spots']
```

Figure 5. Sample of generated prompts

7.3 Model Fine-Tuning

The CLIP model used in our training is the 'ViT-B/32' model, chosen for its balance of performance and computational efficiency. Certain layers of the model are unfrozen to improve performance for our task:

Visual encoder: The first convolutional layer, final normalization layer, and the last N transformer blocks, which are unfrozen gradually during training.

Text encoder: remains frozen. This is done because we are not retraining CLIP to understand new language but training the vision side to align with our prompts better.

7.4 Training and Validation

The model is trained for 30 epochs with a batch size of 64, and an initial learning rate of 1e-5, optimized using the AdamW optimizer with betas = (0.9, 0.99) and a weight decay of 1e-5 to mitigate overfitting. Learning rate decay is managed using a CosineAnnealingLR scheduler, allowing for smooth reduction over epochs.

During training, the model is set to .train() mode, and for each batch:

1. Images and labels are loaded onto the device.
2. Visual embeddings are generated using CLIP's encode_image method.
3. Similarity scores are computed between the image embedding and each prompt embedding for both classes.
4. For each class, the average similarity score is calculated and combined into a two-element logit vector.
5. Cross-entropy loss is computed between logits and ground-truth labels and backpropagated.
6. Performance metrics (accuracy, precision, recall, F1 score, and loss) are recorded per epoch.

Validation is performed at the end of each epoch with .eval() mode and torch.no_grad() to disable gradient tracking. The validation loop mirrors the training loop, computing embeddings and similarity scores to produce predictions. If the current validation loss is lower than previously recorded values, the model checkpoint is saved. The learning rate is updated at each epoch using the scheduler.

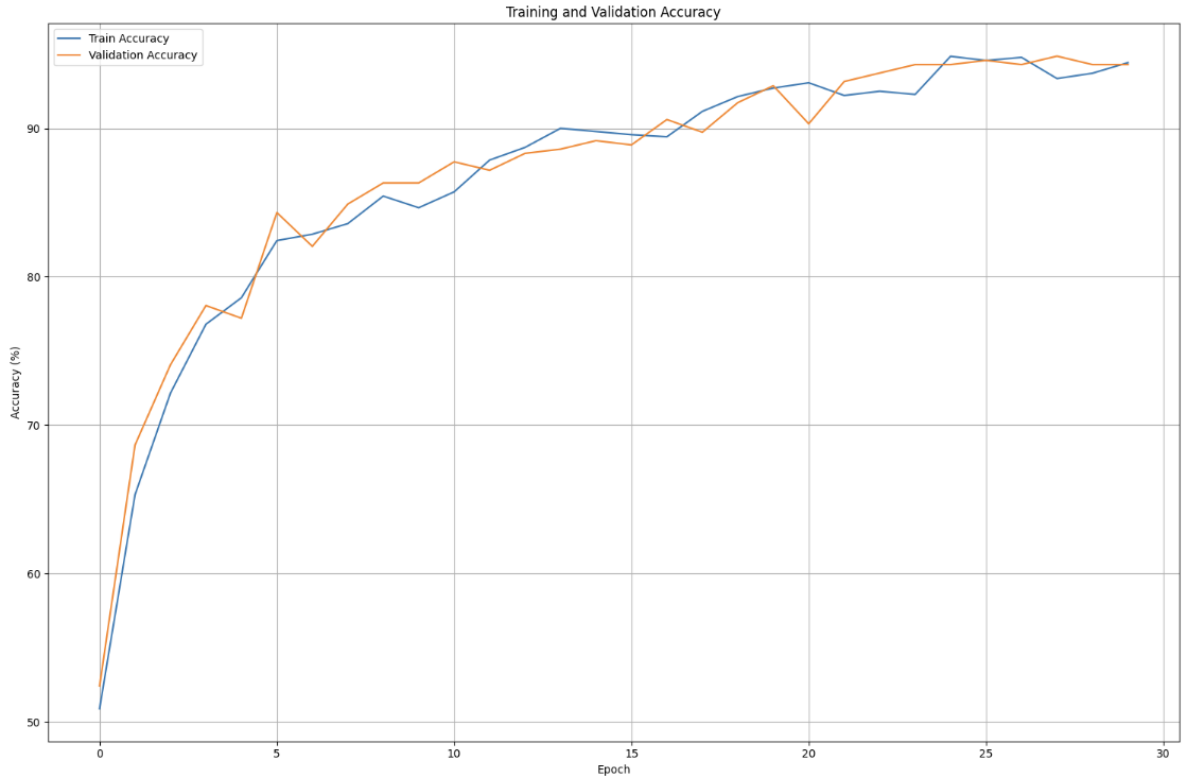

Figure 6. Training and validation accuracy over training loops

7.5 Model Testing

The evaluate_dataset function is used to evaluate the trained model on the other datasets. A DataLoader is initialized without shuffling to maintain consistent sample order. With the model in evaluation mode, image embeddings are generated and compared with prompt embeddings via dot product similarity. The average similarities per class are computed, and predictions are made based on the higher score.

Evaluation metrics such as accuracy, precision, recall, and F1 score, are computed using weighted averaging. A detailed classification report and confusion matrix are generated and saved for each test set, providing insight into performance and misclassification patterns.

7.6 Performance Metrics

Model performance is assessed through a range of quantitative and qualitative metrics:

An accuracy plot showing the classification results for each dataset is plotted, showing how well the model performed across the different testing sets.

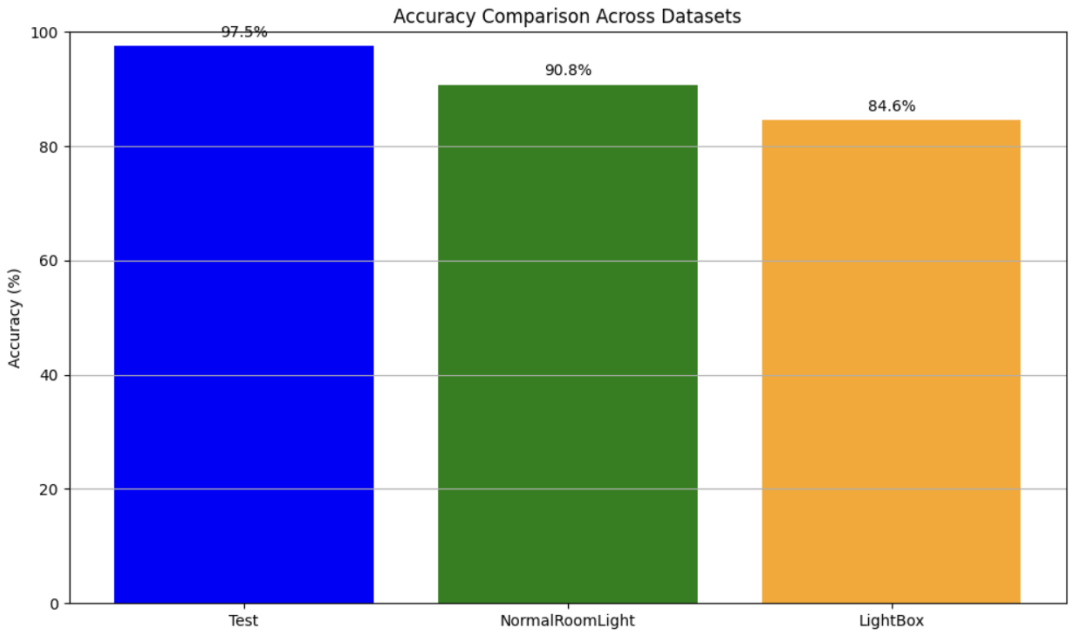

Figure 7. Sample dataset accuracy plot

A classification report is generated for each of the testing sets, containing precision, recall, and F1 score values, which highlight the performance of the model on each of the seed classes, revealing performance imbalances or biases.

A confusion matrix is also generated visualizing the actual classification decisions the model made, which gives insights into where it may be failing to differentiate between the two seed classes.

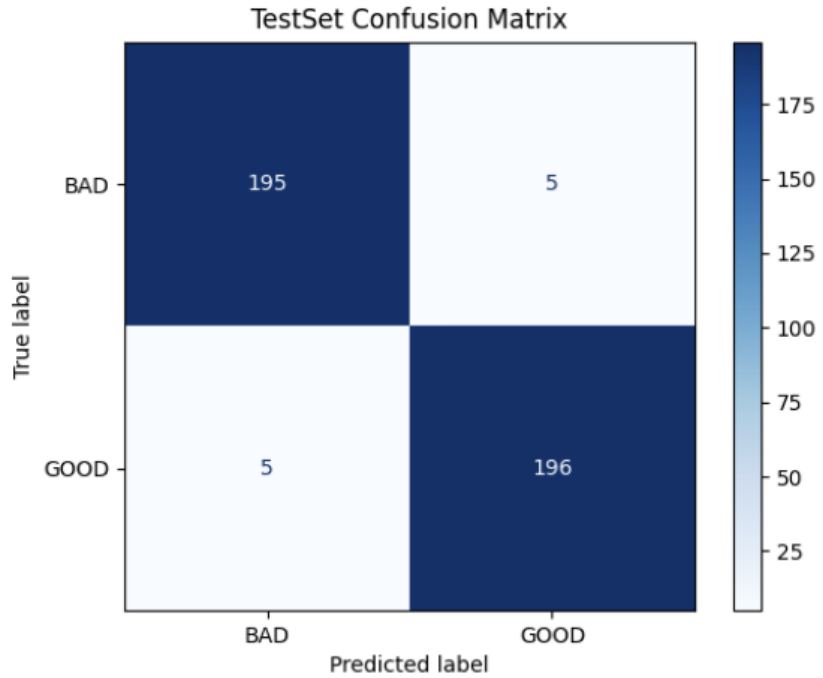

Figure 8. Confusion matrix

Score-CAM heatmaps are generated for a small number of samples from each class, showing the image regions that contributed most to the model's decision.

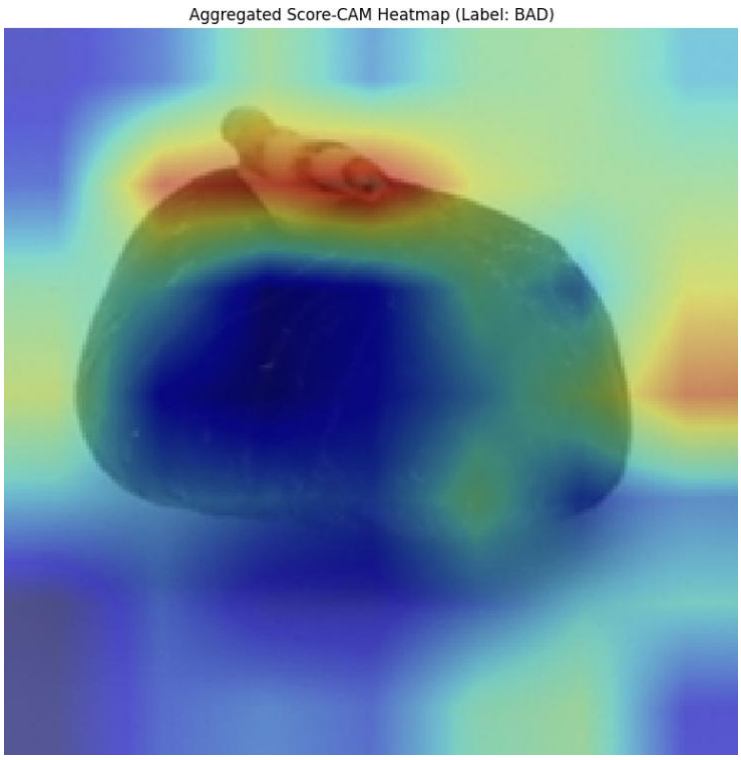

Figure 9. Heatmap for an unhealthy seed sample

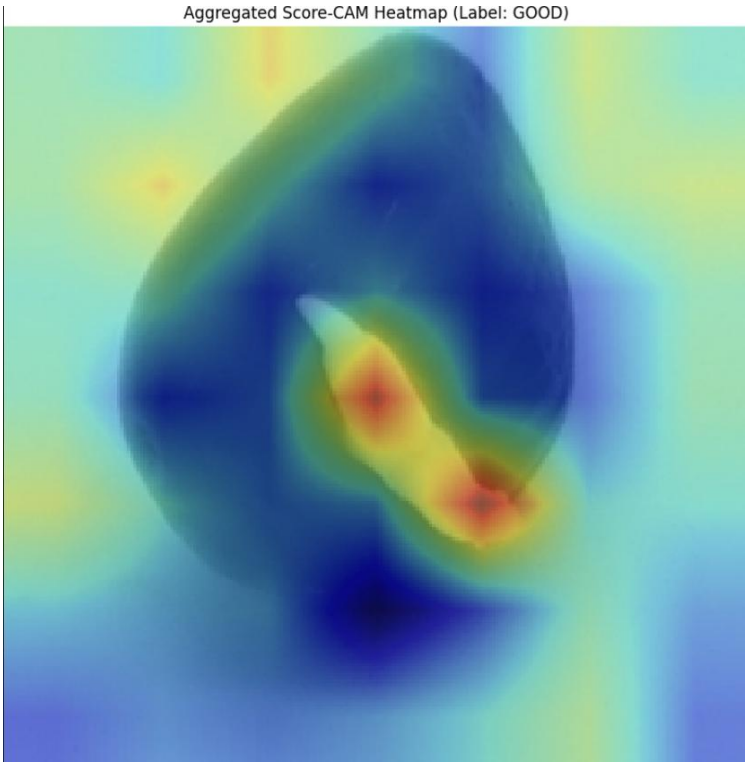

Figure 10. Heatmap for a healthy seed sample

For each prompt, key metrics such as precision, recall, and F1 score are plotted, helping identify strong and weak textual descriptors.

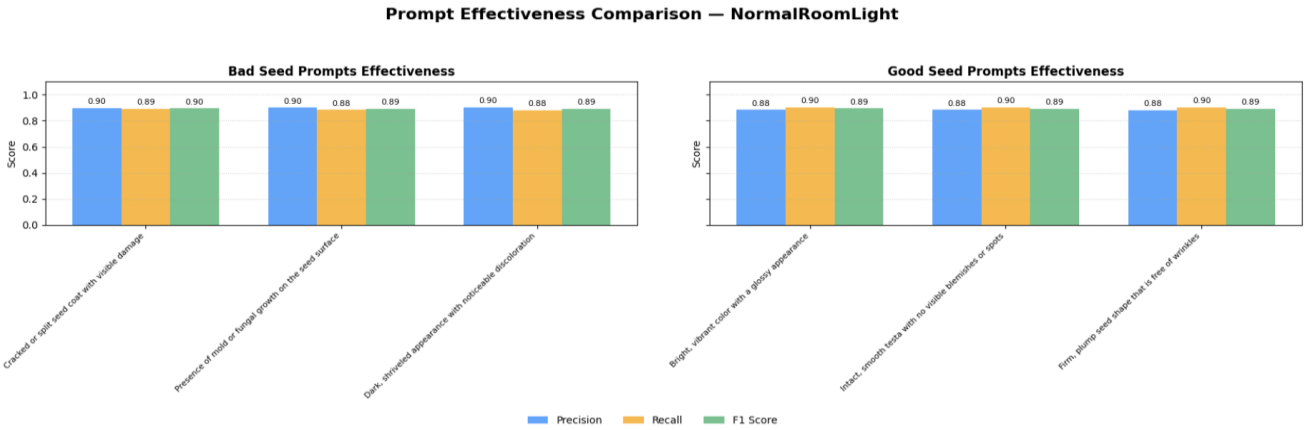

Figure 11. Prompt effectiveness plot

The average similarity for each prompt across each of the two classes is plotted, to show which prompts were the most influential in the classification process, aiding in further refinement of underperforming prompts.

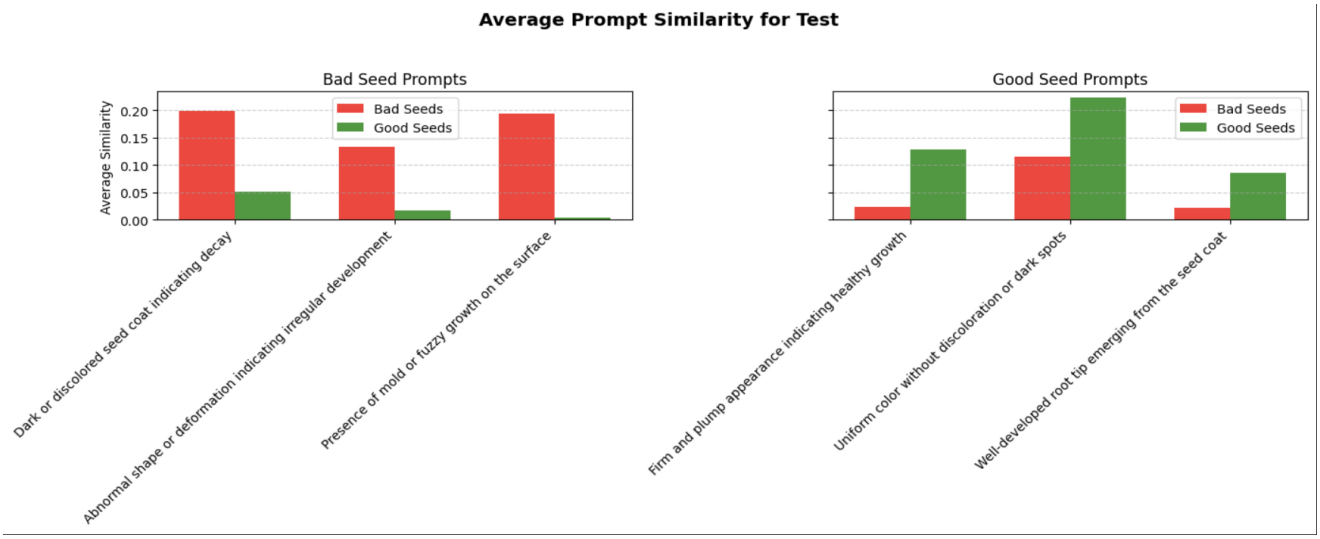

Figure 12. Prompt similarity plot

8 Evaluation

To evaluate the model's performance, we used different sets of prompts that differed in length and complexity. We also assessed the impact of augmentation strategies throughout the training process.

8.1 CLIP Zero-Shot Testing

8.1.1 Zero-Shot with CLIP Recommended Prompts

This first test was conducted using the prompt format recommended in the CLIP paper [5], which is “a photo of a [CLASS]”. For our binary classification task, we used the class-specific prompts:

- Unhealthy seed prompt: “a photo of a bad germinated oil palm seed”
- Healthy seed prompt: “a photo of a good germinated oil palm seed”.

No image augmentations were applied to the dataset, and the CLIP model was used in its pre-trained state without any fine-tuning. Predictions were based solely on the similarity between image embeddings and text embeddings generated from the prompts.

In this experiment, the unhealthy seeds were used as the first class and as a result the model identified all the seeds as “bad” across all the datasets.

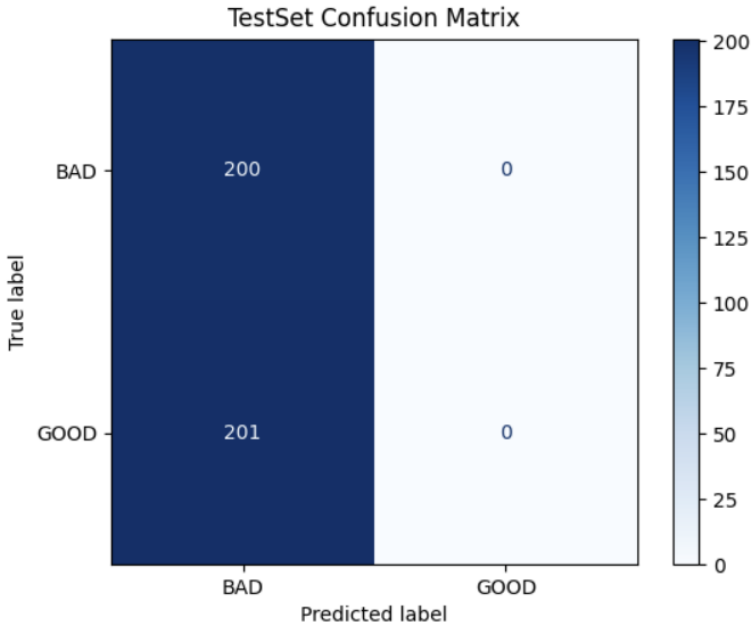

Figure 13. Confusion matrix for test set for zero-shot experiment

Although CLIP is known for its zero-shot classification capabilities, our dataset of germinated oil palm seeds is too specific for an untrained CLIP model to perform well on. As a result, a more comprehensive training strategy is required to obtain satisfactory results.

8.1.2 Zero-Shot with Image Augmentations

This experiment was done using the same class prompts as the previous test, with the key difference being the application of a set of image augmentations to all three datasets. The goal was to check if applying these augmentations would help in increasing model generalization.

The specific image augmentations applied for this test were:

- **Resize:** Resizes the images to 256x256 pixels using bicubic interpolation
- **RandomResizedCrop**: Randomly crops a portion of the original image
- **RandomRotation**: Rotates the image within ±30 degrees
- **RandomVerticalFlip**: 30% chance of vertically flipping the image
- **RandomHorizontalFlip**: 50% chance of horizontally flipping the image
- **ColorJitter**: Randomly adjusts brightness, contrast, saturation by up to ±40% and hue by ±0.1 in the image
- **RandomAffine**: Applies a combination of rotation (±15 degrees), translation (up to 10% of image size), scaling (90–110%), and shearing (±10 degrees)
- **RandomErasing**: Randomly removes rectangular patches covering 2-20% of the image with a 30% probability
- **GaussianBlur**: Applies a Gaussian blur with a kernel size of 3 and a sigma between 0.1 and 2.0

These augmentations were designed to increase data diversity, reduce overfitting, and prepare the model for real-world challenges, such as variations in lighting, orientation, and image imperfections. By applying these transformations, the experiment aimed to improve the CLIP model's generalization, particularly for the NormalRoomLight and Lightbox datasets, which present distinct imaging conditions compared to the test set.

After the datasets were loaded, and had the specified augmentations applied to them, the model was tested on them without any training or fine-tuning.

Dataset	Test	NormalRoomLight	Lightbox
Accuracy (%)	50.1	59.7	54.5

Table 1. Accuracy values with augmentations applied

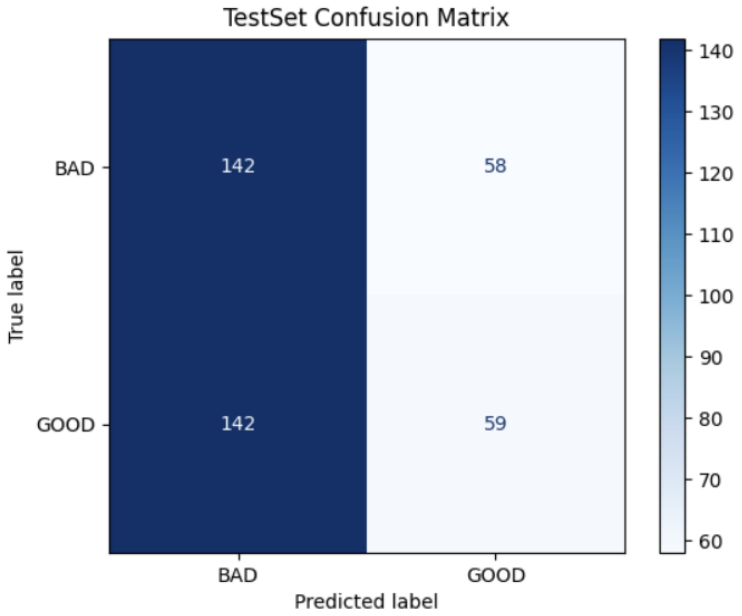

Figure 14. Confusion matrix for test set

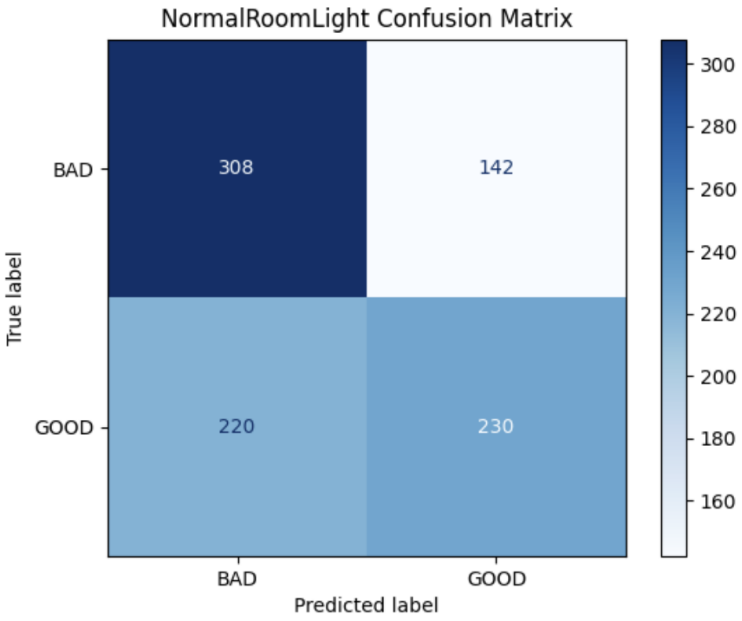

Figure 15. Confusion matrix for NormalRoomLight set

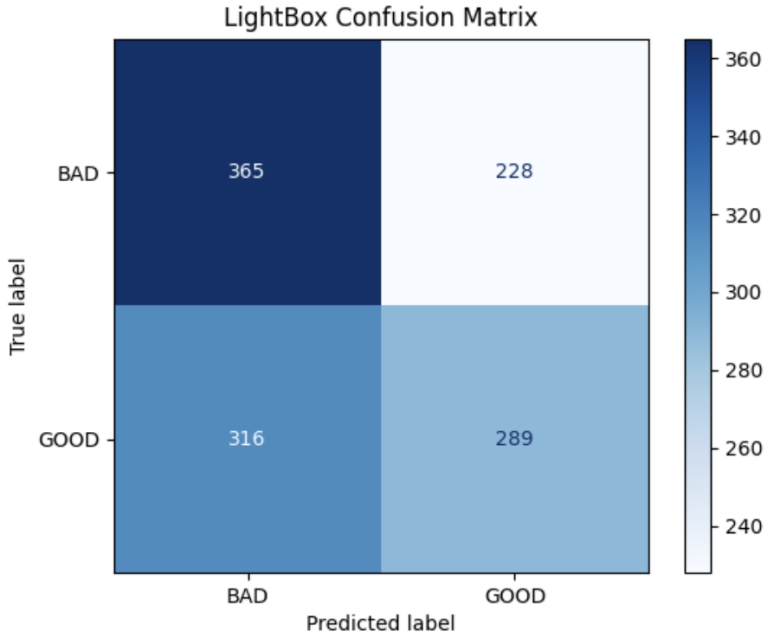

Figure 16. Confusion matrix for Lightbox set

As shown in the results above, applying augmentations to the images allowed the model to perform better due to the visual variations. In particular, performance on the NormalRoomLight and Lightbox sets improved due to the model's increased exposure to lighting diversity, orientation shifts, and minor distortions introduced by the augmentations. These enhancements suggest that even without fine-tuning, well-designed augmentations can help the model better handle real-world variability and maintain consistent performance across different imaging conditions.

8.2 CLIP Model Training with ChatGPT Prompts

8.2.1 Training Using Simple Class Prompts

In this experiment, a set of relatively simple prompts were generated using the ChatGPT API call, three for each seed class. These prompts were short and focused on easily identifiable features like color and surface texture.

The prompts generated for each seed class were as follows:

- Unhealthy seed prompts:
  - "Dark or discolored seed coat indicating decay"
  - "Root sprout appears underdeveloped or has brown discoloration"
  - "Presence of mold or fuzzy growth on the surface"
- Healthy seed prompts:
  - "Firm and plump appearance indicating healthy growth"
  - "Uniform color without discoloration or dark spots"
  - "Healthy white root tip emerging from the seed"

This test was done with the parameters specified below, and with the same image augmentations used before.


```
"model": "ViT-B/32",
"num_descriptors": 3,
"batch_size": 64,
"num_epochs": 30,
"learning_rate": 1e-05,
"optimizer": "AdamW",
"scheduler": "CosineAnnealingLR",
"initial_unfrozen_layers": 6,
"target_unfrozen_layers": 10,
"weight_decay": 1e-05,
"betas": [
    0.9,
    0.99
],
"data_augmentation": [
    "Resize", "RandomResizedCrop", "RandomRotation",
    "RandomHorizontalFlip", "RandomVerticalFlip", "ColorJitter",
    "RandomAffine", "RandomErasing","GaussianBlur"
```


Figure 17. Parameters used for training

Dataset	Test	NormalRoomLight	Lightbox
Accuracy (%)	97.8	93.6	86.1

Table 2. Accuracy values across datasets after training

The accuracy values in Table 2 show that even these simple prompts enabled the model to perform well across the different datasets. The model showed especially good performance in the NormalRoomLight set, the Lightbox result values were slightly lower, but still strong, indicating that the model could effectively apply the prompt information across the different lighting conditions.

Seed Class	Precision	Recall	F1 Score
Unhealthy	0.938	0.933	0.935
Healthy	0.934	0.938	0.936

Table 3. Results for NormalRoomLight set

Seed Class	Precision	Recall	F1 Score
Unhealthy	0.866	0.853	0.860
Healthy	0.858	0.870	0.864

Table 4. Results for Lightbox set

The results in Tables 3 and 4 demonstrate that our model performs well across both the NormalRoomLight and Lightbox datasets, achieving high precision, recall, and F1 scores in both cases. However, a noticeable performance gap exists between the two datasets.

Performance on the NormalRoomLight dataset (Table 3) was particularly strong and balanced, with F1 scores exceeding 0.93 for both classes. This indicates that simple prompts effectively guided the model towards identifying basic but meaningful seed features like color uniformity or mold presence.

On the Lightbox dataset (Table 4), there was a slight drop in the model's performance, with F1 scores decreasing to approximately 0.86 for both classes. This decline could be attributed to the different lighting conditions in the Lightbox setup, which differ from the seed images present in the training data, even after augmentations. As a result, the model may have failed to generalize certain visual features.

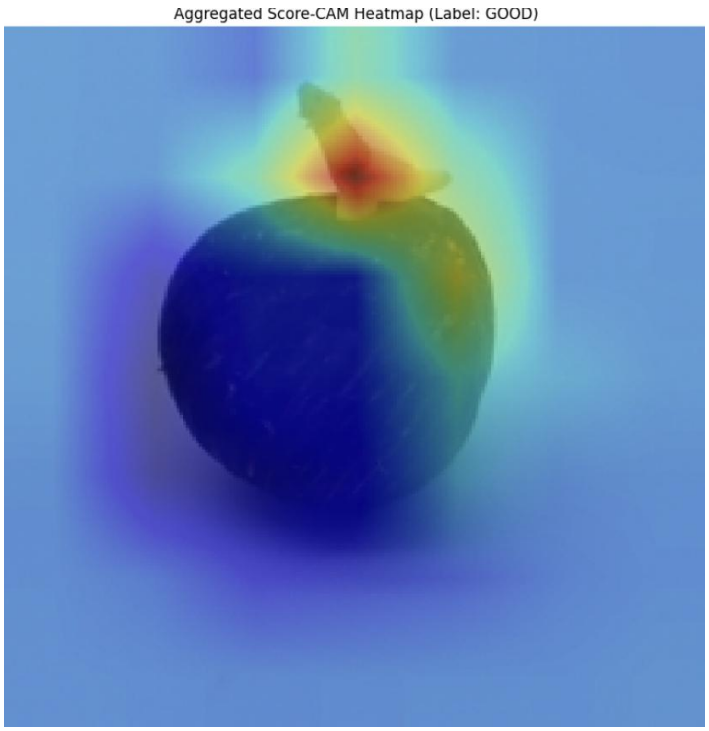

Figure 18. Healthy seed heatmap

In addition to the improved classification accuracy across the datasets, the model was also able to correctly highlight meaningful regions within the input images. As illustrated in Figure 18, the heatmap corresponding to a correctly identified healthy seed highlights a key visual feature, in the form of a healthy root sprout, which is an indicator strongly associated with a good seed. This suggests that the model is not only making correct decisions but also focusing on relevant regions that align with human-understandable attributes of the target class.

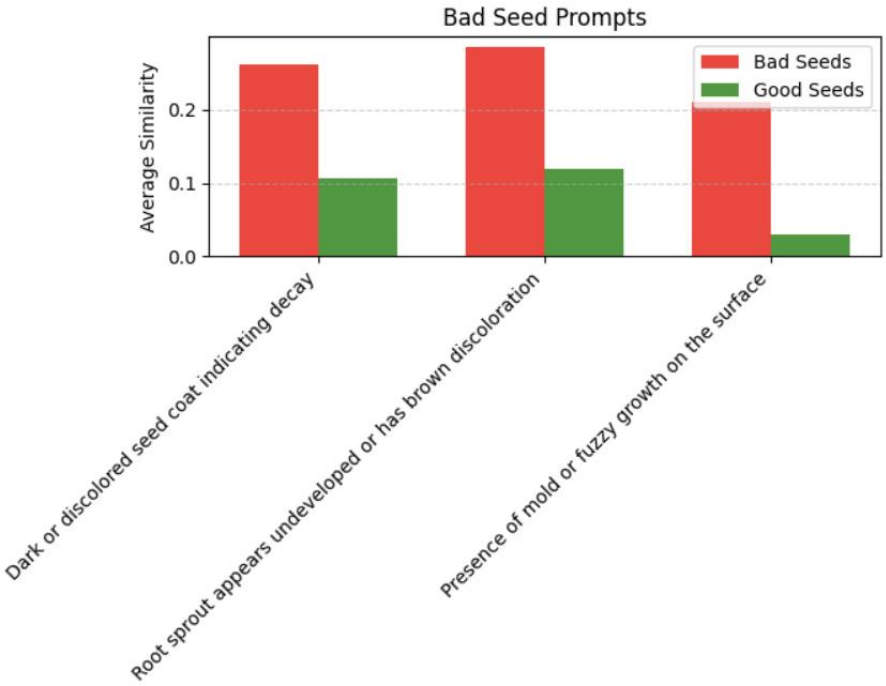

Figure 19. Unhealthy seed prompt similarity plot

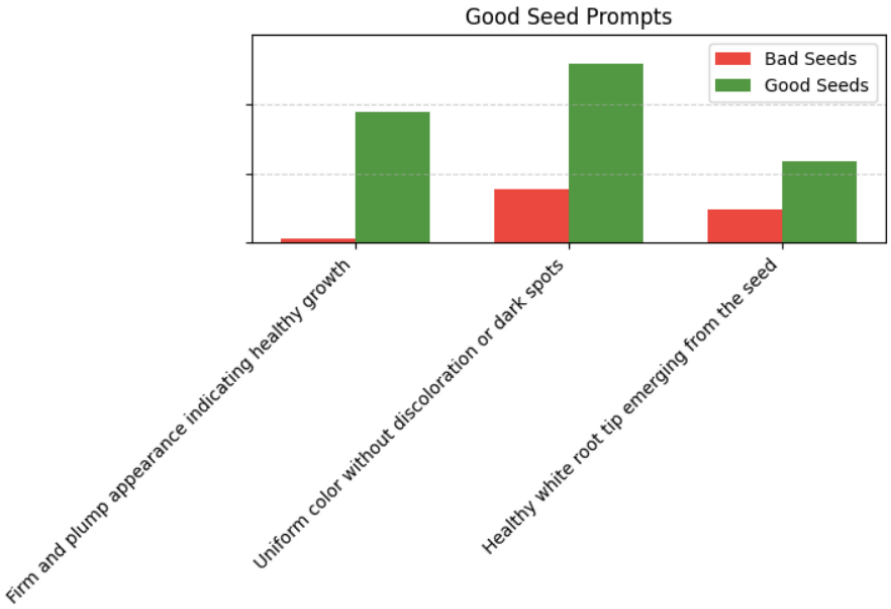

Figure 20. Healthy seed prompt similarity plot

The prompt similarity plot shows that effective prompts consistently produced higher average similarity scores with their corresponding class than with the opposite class. For instance, prompts like “Prescence of mold or fuzzy growth on the surface” for the unhealthy seeds and “Firm and plump appearance indicating healthy growth” for healthy seeds, were clearly distinguished by the model to the target class, indicating its ability to associate these descriptors with the correct seed images.

8.2.2 Training with More Complex Prompts

For this test the API call to ChatGPT was changed so that instead of a few general words, the generated prompts were more complex and specific to each class of seeds. This experiment was done to test if using more concise prompts would help the model to perform better, if it is able to capture those nuanced seed characteristics.

The specific prompts generated by ChatGPT for this test were:

- Unhealthy seed prompts:
  - "Presence of a black or dark brown discoloration on the embryo axis, indicating potential fungal infection or necrosis"
  - "Appearance of a wrinkled or shriveled seed coat surface, suggesting dehydration or poor seed development"
  - "Visible growth of mold or fungal structures on the seed coat, characterized by white or grey filamentous patches"
- Healthy seed prompts:
  - "The plumule is a robust and upward-pointing shoot with a vibrant green color, indicating healthy growth"
  - "The radicle is visibly elongated and white, emerging clearly from the seed, showing active root development"
  - "The seed coat exhibits a uniform dark brown color without any patches, indicating intact protective layers"

Dataset	Test	NormalRoomLight	Lightbox
Accuracy (%)	97.0	91.8	85.2

Table 5. Accuracy using more complex prompts

The use of more complex and descriptive prompts, such as references to specific seed properties (e.g. plumule, radicle) had an influence on the model's performance.

Overall accuracy decreased slightly compared to the simpler prompt experiment, suggesting that while the detailed prompts capture more class-specific features, they may have introduced complexity that was harder for the model to consistently align with visual features in certain conditions.

Seed Class	Precision	Recall	F1 Score
Unhealthy	0.950	0.882	0.915
Healthy	0.890	0.953	0.921

Table 6. Results for NormalRoomLight set

In the NormalRoomLight set (Table 6), the model achieved strong performance with high F1 scores across both classes. The high recall for healthy seeds (0.953) suggests that the model was particularly effective at identifying healthy seed when guided by prompts that described specific positive indicators such as "elongated radicle" and "uniform dark brown seed coat".

Seed Class	Precision	Recall	F1 Score
Unhealthy	0.926	0.762	0.836
Healthy	0.801	0.940	0.865

Table 7. Results for Lightbox set

In contrast, performance on the Lightbox set was more uneven. While the healthy seed class retained a high recall, the bad class recall dropped significantly to 0.762, indicating that the model struggled to consistently identify unhealthy seeds for this dataset. One possible explanation is that the detailed prompts visual properties (e.g. “filamentous patches”, “shrivelled surfaces”) that were either less visually prominent or harder to distinguish in the Lightbox images.

Although the prompts used in this experiment were more detailed, this did not translate into improved model performance or generalizability across datasets. In fact, the added complexity may have introduced overly specific expectations that the model failed to consistently match, particularly under varying lighting conditions.

This suggests that highly detailed prompts, while potentially beneficial in controlled settings, can reduce robustness when visual features are less pronounced or harder to detect. Simpler prompts, although less descriptive, may offer general alignment with visual data and more stable performance across diverse scenarios. Highlighting the need for balance between prompt clarity and specificity for reliable multi-modal classification.

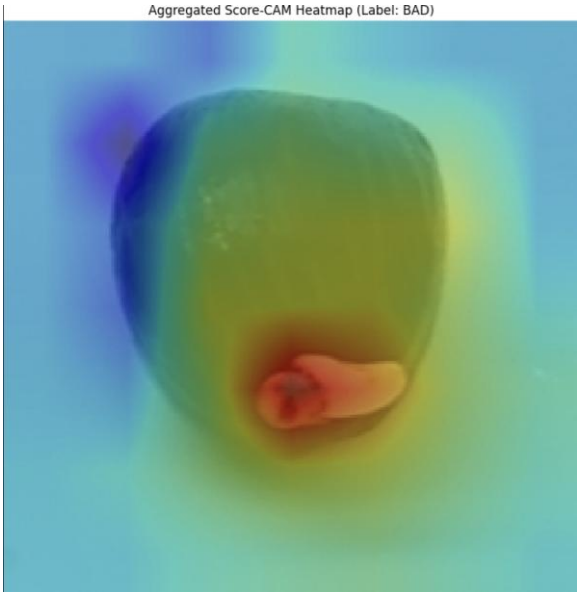

Figure 21. Unhealthy seed Score-CAM

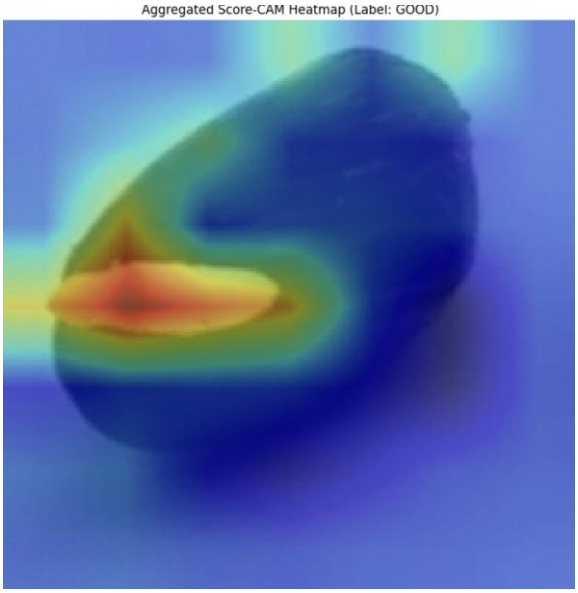

Figure 22. Healthy seed Score-CAM

Even with the usage of the more complex prompts, the model was still able to highlight areas of high interest in the seed images, such as the discolored sprout as shown in Figure 21, and the healthy one in Figure 22.

Unhealthy seed prompt	Average bad similarity	Average good similarity
Presence of a black or dark brown discoloration on the embryo axis, indicating potential fungal infection or necrosis	0.375	-0.359
Appearance of a wrinkled or shriveled seed coat surface, suggesting dehydration or poor seed development	0.240	-0.223
Visible growth of mold or fungal structures on the seed coat, characterized by white or grey filamentous patches	0.330	-0.312

Table 8. Average similarity values for bad seed prompts

Healthy seed prompt	Average good similarity	Average bad similarity
The plumule is a robust and upward-pointing shoot with a vibrant green color, indicating healthy growth	0.334	-0.337
The radicle is visibly elongated and white, emerging clearly from the seed, showing active root development	0.163	-0.131
The seed coat exhibits a uniform dark brown color without any patches, indicating intact protective layers	0.024	-0.001

Table 9. Average similarity values for healthy seed prompts

The average similarity values in the tables offer valuable insights into the relative effectiveness of each prompt. These scores help explain why some prompts lead to stronger classification performance than others. For example, the healthy seed prompt “The seed coat exhibits a uniform dark brown color without any patches, indicating intact protective layers” had a very low average similarity with the good seed class (0.024) compared to the other prompts used. One possible explanation is that the phrase “indicating intact protective layers” may be too verbose or abstract for the model to associate directly with visual features, reducing the prompt’s effectiveness.

In contrast, the healthy seed prompt “The plumule is a robust and upward-pointing shoot with a vibrant green color, indicating healthy growth”, achieved the highest average similarity with the good seed class (0.334), while maintaining a low similarity with the bad seed class (-0.337). This suggests that the model was able to easily distinguish that visual feature when it was present in the seed images, making it an effective descriptor for classification.

8.2.3 Prompt Refinement Process

Based on the results of the previous run, we conducted an additional experiment to replace the ineffective healthy seed descriptor with a more effective one, aiming to improve the model's ability to distinguish between the class prompts.

This time we used the API call to ChatGPT to generate only one healthy seed descriptor to replace the weak one used in the previous test.

The prompt generated was the following: "The seed coat appears smooth and slightly glossy with a uniform texture and no irregularities"

The updated set of prompts used in this test were as follows:

- Unhealthy seed prompts:
    - "Presence of a black or dark brown discoloration on the embryo axis, indicating potential fungal infection or necrosis"
    - "Appearance of a wrinkled or shriveled seed coat surface, suggesting dehydration or poor seed development"
    - "Visible growth of mold or fungal structures on the seed coat, characterized by white or grey filamentous patches"
- Healthy seed prompts:
    - "The plumule is a robust and upward-pointing shoot with a vibrant green color, indicating healthy growth"
    - "The radicle is visibly elongated and white, emerging clearly from the seed, showing active root development"
    - "The seed coat appears smooth and slightly glossy with a uniform texture and no irregularities"

The model was then trained using the same procedure as the previous experiment.

Healthy seed prompt	Average good similarity	Average bad similarity
The plumule is a robust and upward-pointing shoot with a vibrant green color, indicating healthy growth	0.349	-0.324
The radicle is visibly elongated and white, emerging clearly from the seed, showing active root development	0.151	-0.139
The seed coat appears smooth and slightly glossy with a uniform texture and no irregularities	0.125	-0.113

Table 10. Healthy seed similarities with updated prompt

The results showed that the new prompt used improved the model's ability at separating between the two target classes. Its simpler and more direct language made it easier for the model to interpret and use effectively in classification.

This experiment highlights CLIP's sensitivity to prompt wording and emphasizes the importance of prompt quality and consistency in fine-tuning. To ensure robust performance, it is essential to carefully curate prompts that clearly describe class-specific features.

CLIP's similarity scores could serve as part of an automated prompt selection pipeline [18]. By evaluating an initial pool of candidate prompts and selecting only the ones that yield high similarity with the correct class, and low similarity with the opposite one, prompt engineering can be streamlined [20]. This approach prioritizes descriptors that the model responds to most reliably, ultimately enhancing performance and reducing manual prompt tuning.

9 Summary and Reflections

This implementation explored the use of a CLIP-based model for classifying seed quality by leveraging image-text similarity, with a focus on prompt effectiveness and model interpretability. Through evaluation across multiple datasets, we found that aligning textual prompts with visual seed characteristics played a critical role in achieving strong classification performance. Furthermore, heatmap visualizations provided insights into the model's attention mechanisms, demonstrating its ability to focus on semantically relevant regions of the image.

To streamline the prompt engineering process, we leveraged ChatGPT to generate descriptive prompts for both good and bad seed categories. This approach reduced the need for manual prompt design and allowed for consistent exploration of textual cues. The results show that the prompts crafted by ChatGPT often aligned effectively with the corresponding seed images, supporting the feasibility of using language models to automate prompt creation.

Overall, these findings emphasize the importance of prompt quality in vision-language models, particularly in domains where subtle visual differences carry significant semantic meaning. While CLIP has demonstrated impressive zero-shot capabilities, its performance remains closely tied to prompt quality [17] [19]. This highlights a broader challenge in prompt-based learning, which is achieving consistency and robustness without relying heavily on manual prompt tuning.

References

1. Santos, Luís & Neves Dos Santos, Filipe & Moura Oliveira, Paulo & Shinde, Pranjali. (2020). Deep Learning Applications in Agriculture: A Short Review. 10.1007/978-3-030-35990-4_12.
2. de Medeiros, A. D., Capobiango, N. P., da Silva, J. M., da Silva, L. J., da Silva, C. B., & dos Santos Dias, D. C. F. (2020). Interactive machine learning for soybean seed and seedling quality classification. *Scientific reports*, *10*(1), 11267.
3. TU, K. L., LI, L. J., YANG, L. M., WANG, J. H., & Qun, S. U. N. (2018). Selection for high quality pepper seeds by machine vision and classifiers. *Journal of Integrative Agriculture*, *17*(9), 1999-2006.
4. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020). Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. *ArXiv:1910.01279 [Cs]*. https://arxiv.org/abs/1910.01279
5. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *ArXiv:2103.00020 [Cs]*. https://arxiv.org/abs/2103.00020
6. Suh, Hyun & IJsselmuiden, Joris & Hofstee, J.W. & Van Henten, E.J.. (2018). Transfer learning for the classification of sugar beet and volunteer potato under field conditions. Biosystems Engineering. 174. 50-65. 10.1016/j.biosystemseng.2018.06.017.
7. Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to Prompt for Vision-Language Models. *ArXiv:2109.01134 [Cs]*. https://arxiv.org/abs/2109.01134
8. Perez, L., & Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *ArXiv:1712.04621 [Cs]*. https://arxiv.org/abs/1712.04621
9. Howard, J., & Ruder, S. (n.d.). *Universal Language Model Fine-tuning for Text Classification*. https://arxiv.org/pdf/1801.06146
10. Dong, X., Bao, J., Zhang, T., Chen, D., Gu, S., Zhang, W., Yuan, L., Chen, D., Wen, F., & Yu, N. (2022). *CLIP Itself is a Strong Fine-tuner: Achieving 85.7% and 88.0% Top-1 Accuracy with ViT-B and ViT-L on ImageNet*. ArXiv.org. https://arxiv.org/abs/2212.06138
11. Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, *147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016
12. Loddo, A., Loddo, M., & Di Ruberto, C. (2021). A novel deep learning based approach for seed image classification and retrieval. *Computers and Electronics in Agriculture*, *187*, 106269. https://doi.org/10.1016/j.compag.2021.106269
13. Corley, R. H. V., & Tinker, P. B. (2015). *The Oil Palm*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118953297

14. Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., & Beyer, L. (2021). How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. *ArXiv:2106.10270 [Cs]*. https://arxiv.org/abs/2106.10270
15. Madasu, A., Lal, V., & Howard, P. (2025). *Quantifying Interpretability in CLIP Models with Concept Consistency*. ArXiv.org. https://arxiv.org/abs/2503.11103
16. Storås, A. M., Ole Emil Andersen, Lockhart, S., Thielemann, R., Filip Gnesin, Vajira Thambawita, Hicks, S. A., Kanters, J. K., Strümke, I., Halvorsen, P., & Riegler, M. A. (2023). Usefulness of Heat Map Explanations for Deep-Learning-Based Electrocardiogram Analysis. *Diagnostics*, *13*(14), 2345–2345. https://doi.org/10.3390/diagnostics13142345
17. Li, A., Liu, Z., Li, X., Zhang, J., Wang, P., & Wang, H. (2025). *Modeling Variants of Prompts for Vision-Language Models*. ArXiv.org. https://arxiv.org/abs/2503.08229
18. Metzen, J. H., Piyapat Saranrittichai, & Chaithanya Kumar Mummadi. (2024). *AutoCLIP: Auto-tuning Zero-Shot Classifiers for Vision-Language Models*. Transactions on Machine Learning Research. https://openreview.net/forum?id=gVNyEVKjqf
19. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., & Zou, J. (2023, March 23). *When and why vision-language models behave like bags-of-words, and what to do about it?* ArXiv.org. https://doi.org/10.48550/arXiv.2210.01936
20. Lu, Y., Liu, J., Zhang, Y., Liu, Y., & Tian, X. (n.d.). *Prompt Distribution Learning*. Retrieved May 4, 2025, from https://openaccess.thecvf.com/content/CVPR2022/papers/Lu_Prompt_Distribution_Learning_CVPR_2022_paper.pdf