COSE474-2024F: Final Project Report
"**Adapting CLIP for Personal Color Analysis**"

Hanjun Park

1 Introduction

Motivation. Personal color analysis is a subjective yet impactful approach that helps individuals determine which color palettes suit their physical appearance best. Personal colors are typically divided into four categories—spring, summer, fall, and winter—based on physical characteristics. Traditionally, personal color analysis requires a trained professional, making it inaccessible to many. The motivation behind this project is to democratize personal color analysis by making it accessible to anyone with an AI-driven solution.

Problem Definition. The primary problem we address is whether a pre-trained model, specifically CLIP (Contrastive Language-Image Pretraining), can accurately classify personal colors into these four categories based solely on visual features. We also explore what modifications or tuning methods are needed to enhance CLIP's classification performance.

Contribution. The study explores several learning strategies, including zero-shot, fine-tuning, and prompt-based tuning, with a focus on identifying the most reliable method for personal color classification[4]. This research aims to democratize personal color analysis, making it accessible to a broader audience without requiring expert intervention.

2 Related Works

CLIP. CLIP is a powerful pre-trained vision-language model developed by Radford et al. [4]. It learns to associate images and text in a shared embedding space through contrastive learning. This property allows CLIP to perform well in zero-shot classification tasks, such as personal color classification, without extensive retraining.

Zero-shot Learning. Zero-shot learning has gained significant attention due to its efficiency in leveraging pre-trained models for new tasks without requiring task-specific training data. In this context, CLIP's inherent ability to align visual and textual representations makes it particularly effective for zero-shot classification scenarios, including personal color categorization.

Fine-tuning. Fine-tuning is a common method to adapt pre-trained models for specific tasks by updating their weights. Fully fine-tuning all parameters of large models like CLIP can significantly improve performance for task-specific applications but is computationally intensive. To address this, LoRA adapts the model by introducing low-rank matrices, thereby reducing computational overhead while retaining adaptation effectiveness [1].

Prompt Tuning. Prompt tuning is an approach to adapt pre-trained models by adding contextual prompts to their inputs, allowing them to perform better on specific tasks with minimal changes to the model's parameters. Prompt tuning techniques, such as CoCoOp (Conditional Context Optimization) and P-Tuning v2, modify the input prompts fed into CLIP to enhance its classification performance for specific categories [5, 3]. These methods are particularly useful for adapting large pre-trained models like CLIP without incurring the computational costs of full fine-tuning.

3 Methods

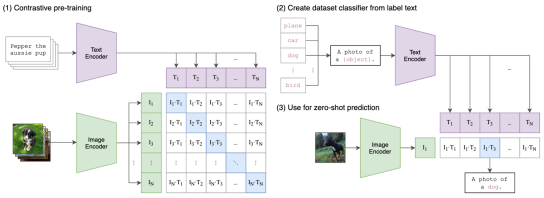

Figure 1: The architecture of CLIP

This section describes the specific methods employed in this project. All methods build upon CLIP. Figure 1 illustrates the core architecture of CLIP, which serves as the foundation for all tuning approaches used in this study. The significance of this study lies in its systematic exploration of multiple adaptation methods to evaluate CLIP's potential for domain-specific tasks like personal color classification.

3.1 Fine-tuning

Fine-tuning adapts the pre-trained CLIP model to the specific domain of personal color classification by updating its weights. Fully fine-tuning allows the model to learn domain-specific features, improving classification accuracy, though

it is computationally intensive and risks overfitting on smaller datasets. For this experiment, we trained the model for 10 epochs, observing the highest performance during this period. This approach demonstrated CLIP's ability to adapt to fine-grained, visually subjective tasks through complete parameter updates. The process of fine-tuning is detailed in Algorithm 1.

Algorithm 1 Fine-tuning CLIP

Require: CLIP visual encoder f_V, classifier W_{cls}

for batch $(x, y) \in D$ **do**
  Extract features $v = f_V(x)$
  Compute logits $l = W_{\text{cls}}(v)$
  Update f_V with learning rate η_1
  Update W_{cls} with learning rate η_2
end for
return Fine-tuned f_V, W_{cls}

LoRA-based Fine-tuning. LoRA (Low-Rank Adaptation) is a parameter-efficient technique that fine-tunes large pre-trained models by introducing low-rank matrices, reducing computational costs while retaining adaptation effectiveness. In this project, LoRA was implemented within CLIP's vision transformer blocks by introducing rank-4 projection layers, allowing the model to capture domain-specific features while maintaining computational efficiency. The LoRA-based fine-tuning process is detailed in Algorithm 2.

Algorithm 2 LoRA for CLIP

Require: CLIP visual encoder f_V, LoRA layers $\{L_i\}_{i=1}^{12}$

for batch $(x, y) \in D$ **do**
  Extract features $v = f_V(x)$
  for each L_i **do**
    $v = v + 0.1 \cdot W_{\text{up}}(W_{\text{down}}(v))$
  end for
  Update LoRA parameters using gradient descent
end for
return Optimized $\{L_i\}_{i=1}^{12}$

3.2 Prompt Tuning

CoCoOp. Conditional Context Optimization extends CLIP's flexibility by dynamically generating prompts conditioned on input images. In this study, CoCoOp employed a meta-network to generate context embeddings tailored to personal color categories. These dynamic prompts allowed the model to better align visual and textual features without altering CLIP's core architecture. The CoCoOp tuning procedure is described in Algorithm 3.

Algorithm 3 Conditional Prompt Learning (CoCoOp)

Require: CLIP model, learnable context C, meta-network M_θ

for batch $(x, y) \in D$ **do**
  Generate context features $c = M_\theta(C)$
  Extract image and text features f_I, f_T
  Compute logits $l = 100 \cdot f_I f_T^T$
  Update C, M_θ using gradient descent
end for
return Optimized C, M_θ

P-Tuning v2. Prompt tuning modifies input prompts to better align CLIP's pre-trained embeddings with the task-specific requirements. For this project, P-Tuning v2 introduced learnable prompt embeddings to adapt to the nuanced characteristics of personal color classification. These embeddings were trained using backpropagation while keeping the CLIP parameters frozen, minimizing computational overhead. The P-Tuning v2 prompt tuning procedure is described in Algorithm 4.

Algorithm 4 P-tuning v2 for CLIP

Require: Pre-trained CLIP visual encoder f_V, prompt encoder P_θ

for batch $(x, y) \in D$ **do**
  Extract features $v = f_V(x)$
  Generate prompted features $z = P_\theta(v)$
  Compute cross-entropy loss $\mathcal{L}_{\text{CE}}$
  Update P_θ using gradient descent
end for
return Optimized P_θ

3.3 Evaluation Metrics

We evaluated our methods using two key metrics: accuracy and recall. The overall classification accuracy serves as our primary metric, calculated as:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (1)$$

Additionally, we computed the recall metric for each season category:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (2)$$

These metrics were chosen to assess both the overall performance and the ability to correctly identify specific seasonal color types, which is crucial for the subjective nature of personal color analysis.

4 Experiments

Dataset The dataset used in this study consists of 6,000 labeled images across four personal color

Method	Accuracy	Precision	Recall	F1-Score
Zero-shot	22.31	22.00	24.00	19.00
Fine-tuning	54.22	52.00	52.00	52.00
LoRA	40.34	39.00	40.00	38.00
CoCoOp	21.99	18.00	24.00	19.00
P-Tuning v2	45.78	49.00	46.00	44.00

Table 1: Quantitative results of various tuning methods for personal color classification (all in percentages).

categories: spring, summer, fall, and winter. The images were sourced from the Deep-Armocromia repository [8] and Roboflow. Each category comprises 1,200 training images and 300 test images. The images were preprocessed to match CLIP's input requirements, including resizing to 224×224 pixels.

Computing Resources All experiments were conducted using Google Colab Pro with an NVIDIA A100 GPU. The use of high-performance GPU resources allowed for efficient training and inference.

Experimental Design The experiments were designed to evaluate the effectiveness of various tuning methods for personal color classification:

- **Baseline**: Zero-shot learning using the pre-trained CLIP model.
- **Fine-tuning**: Fully fine-tuning CLIP and Low-Rank Adaptation for the classification task.
- **Prompt Tuning**: Using CoCoOp and P-Tuning v2 for dynamic and learnable prompt tuning.

The primary goal was to identify which method best captures the subtle visual features necessary for accurate personal color classification.

4.1 Quantitative Results

Table 1 summarizes the accuracy, recall, and other metrics for each method. Fine-tuning achieved the highest accuracy (54.2%), followed by P-Tuning v2 (45.8%) and LoRA (40.3%).

Figure 2 visualizes the performance comparison across methods.

4.2 Qualitative Results

Qualitative analysis revealed that fine-tuning produced the most consistent predictions across all color categories, accurately distinguishing subtle differences in skin tones and lighting conditions. P-Tuning v2 also performed well but occasionally misclassified edge cases. Zero-shot learning and CoCoOp showed limited effectiveness, often failing to generalize across diverse inputs.

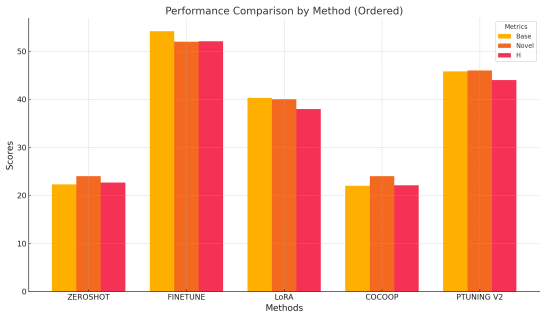

Figure 2: Performance comparison of tuning methods based on accuracy and recall.

4.3 Discussion

Fine-tuning proved highly effective, enabling CLIP to adapt its generalist architecture for fine-grained tasks like personal color classification.[7] Remarkably, the fine-tuned CLIP achieved accuracy comparable to the state-of-the-art FaRL64 (55.4%) on the Armocromia dataset,[8] highlighting its potential for visual feature-based classification without requiring domain-specific pre-training.

In contrast, CoCoOp struggled due to its reliance on dynamic prompt generation,[5] limiting its ability to capture fine-grained features. While P-Tuning v2 and LoRA offered computationally efficient alternatives.[1][3]

This comparison underscores fine-tuning's capability to transform general-purpose models like CLIP into competitive tools for specialized tasks, bridging the gap with state-of-the-art methods.

5 Future Directions

To further enhance the model's capabilities, future efforts could focus on incorporating datasets with broader diversity in skin tones, lighting conditions, and cultural contexts, which would improve the model's generalization and robustness. Additionally, combining the strengths of fine-tuning and prompt-tuning approaches to develop hybrid models could also optimize both performance and computational efficiency.

References

[1] Edward J Hu et al. "Lora: Low-rank adaptation of large language models". In: *arXiv preprint arXiv:2106.09685* (2021).

[2] Xiang Lisa Li and Percy Liang. "Prefix-tuning: Optimizing continuous prompts for generation". In: *arXiv preprint arXiv:2101.00190* (2021).

[3] Xiao Liu et al. "P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks". In: *arXiv preprint arXiv:2110.07602* (2021).

[4] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[5] Kaiyang Zhou et al. "Conditional prompt learning for vision-language models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16816–16825.

[6] Kaiyang Zhou et al. "Learning to prompt for vision-language models". In: *International Journal of Computer Vision* 130.9 (2022), pp. 2337–2348.

[7] Yixuan Wei et al. "Improving clip fine-tuning performance". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 5439–5449.

[8] Lorenzo Stacchio et al. "Deep Armocromia: A Novel Dataset for Face Seasonal Color Analysis and Classification". In: *European Conference on Computer Vision (ECCV) Workshops*. Springer. 2024, pp. 1–16.