Enhancing Robustness to Prompt Variations in Vision Language Models: A Comprehensive Evaluation of CLIP, SigLIP, and CoOp under Noisy Prompts with Ensembling Strategies

Aishah Altamimi
Student IDs: g202501850
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

Supervised by: Dr. Muzammil Behzad
muzammil.behzad@kfupm.edu.sa
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

***Abstract*—Vision Language Models (VLMs) like CLIP and SigLIP are zero-shot VLMs that learn to match images with text prompts rather than class names. This makes the model's behavior affected by the way prompts are structured. This paper builds a reproducible noise benchmark and evaluating model sensitivity to noisy prompts using ensembling methods. The evaluation of the three models CLIP, SigLIP, and CoOp is conducted across five benchmark datasets:Oxford-III Pets, Caltech101, Food101, DTD, and EuroSAT (RGB), each representing a different visual domain. Using clean prompts, noisy prompts with different severity levels, and several test-times ensembling methods, our results show that CLIP is very sensitive to noise when using a single prompt, but test-time ensembling can almost fully recover its clean accuracy. For example, on Oxford Pets, accuracy at the highest noise level rises from 25.2% to 68.8% with the K=5 strategy (four noisy prompts + 1 clean). SigLIP performs better than CLIP on harder datasets such as DTD and EuroSAT, while CoOp stays stable across all noise levels. Finally, we introduce a noise-aware fine-tuning approach that trains a small adapter using corrupted prompts. This further boosts robustness, especially for CLIP, achieving improvements of +30–40% at high noise levels. Overall, the study demonstrates that test-time ensembling and noise-aware fine-tuning improves VLM reliability under prompt corruption.**

***Index Terms*—VLMs, CLIP, SigLIP, CoOp, Prompt Robustness, Noisy Prompts, Contrastive Learning, Zero-Shot Classification, Ensemble Prompting.**

I. Introduction

A. Background and Significance

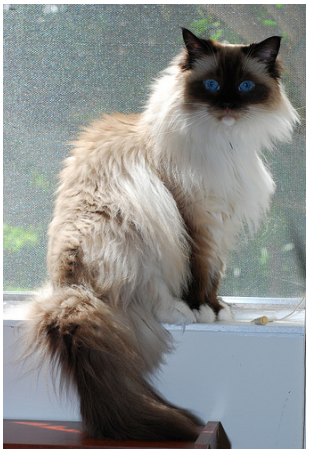
Recent advances in vision–language models (VLMs) include Contrastive Language–Image Pretraining (CLIP), SigLIP (Sigmoid-based Language–Image Pre-Training), and Context Optimization (CoOp). CLIP and SigLIP are zero-shot VLMs that learn to match images with text prompts rather than class names. They encode an image into a vector and a text prompt into another vector, then measure similarity between the two vectors. As this approach depends on text prompt templates such as "a photo of a {class}" instead of the class name. This makes the model's behavior affected by the way prompts are structured [2]. To mitigate this sensitivity, CoOp has been proposed, which replaces CLIP's hand-crafted prompts with learnable continuous tokens optimized for specific tasks [3]. CoOp shows high performance with clean prompts. These models facilitate many tasks such as zero-shot image classification, caption generation, and image–text retrieval [3]. Additionally, VLMs are increasingly used in real-world applications where users interact with systems through natural language prompts—such as image search engines, educational tools (e.g. visual homework helpers), recommendation systems, and e-commerce visual search. In these applications, user queries often contain mistakes, informal phrasing, non-standard spellings, or typographic noise. We observed that these models give high accuracy with clean prompts as illustrated in Figure1 (a); however, they show different behavior under noisy prompts Figure1 (b), and we noticed that Ensembling during test time recovers the accuracy Figure1 (c).

B. Challenges in Current Techniques

Although many applications rely on vision–language models to understand text and connect it with visual information, we still know very little about how stable these models are when the input prompts contain noise or small variations. This makes robustness to prompt changes an essential requirement. In this project, we build a reproducible noise benchmark and assess how sensitive different models are to these variations.By building a reproducible noise benchmark and evaluating sensitivity, this research aims to contribute to the broader domain of robust and trustworthy AI.

C. Problem Statement

Despite vision–language models (VLMs) have improved a lot, studies show that models like CLIP work well with carefully crafted prompt. Which means accuracy can change a lot based on how the text input is phrase [1]. For example, "a photo of a cat" and "an image of the cat" both have same meaning however these variations can result in significant performance differences. Also, adding articles, using synonyms, adding context like "in the wild", "at night", or changing word order. All of these can be treated differently by these



(a) Clean prompt (correct)
Prompt: "a photo of a **birman cat**"
Prediction: birman (✓)

(b) Noisy prompt (misclassified)
Prompt: "a phots of bIrMaN ca t"
Prediction: tabby (✗)

(c) Ensemble of noisy prompts (recovered)
Prompts: averaged over multiple corrupted variants (typo, case, space) of "a photo of a birman cat"
Prediction (averaged logits): birman (✓)

Fig. 1. Illustration of how noisy prompts affect VLMs on the Birman class (Oxford Pets). Clean prompt = correct (a), noisy prompt = incorrect (b), ensembling recovers accuracy (c).

models. To address these issues, this research aims to evaluate the robustness of three VLMs under noisy prompts using ensembling strategy:

• **CLIP** (Contrastive Language-Image Pre-training), relies on manually designed prompts and uses a softmax-based contrastive loss during training [2].

• **SigLIP** (Sigmoid-based Language-Image Pre-training), which also relies on manually crafted prompts but replaces softmax with a sigmoid-based pairwise contrastive loss [2].

• **CoOp** Context Optimization, which is built on top of CLIP and attempts to automate prompt design by learning continuous context tokens [1].

D. Objectives

The main objective of this research is to provide a comprehensive evaluation of various VLMs on their robustness to prompt variation by:

1) Establish baseline for three major vision–language models—CLIP, SigLIP, and CoOp across five benchmark datasets using clean, manually engineered prompts.
2) Systematically evaluate models robustness to noise prompt, including multiple noise types (typos, random casing, spacing noise, emoji noise) and multiple severity levels (0,1,2,3), to quantify how text perturbations affect model predictions.
3) Investigate the effectiveness of prompt ensembling strategies at inference time, comparing configurations such as:
   - K = 1 (single prompt)
   - K = 5 hybrid (clean + noisy prompts)
   - K = 5 all-noise

   to determine whether ensembling can mitigate performance degradation and stabilize predictions.
4) Analyze architectural and training-loss differences among softmax-based contrastive models (CLIP), sigmoid-based pairwise contrastive models (SigLIP), and continuous learned prompt representations (CoOp), in order to understand which paradigm exhibits stronger robustness under prompt variations.
5) Develop and evaluate noise-aware regularization techniques—such as noise-augmented prompt training—to improve resilience beyond baseline zero-shot performance

E. Scope of Study

This study focuses on evaluating the robustness of several vision–language models to noisy text prompts and on developing noise-aware regularization techniques such as noise-augmented prompt training. The evaluation is conducted across five benchmark datasets—Oxford-IIIT Pets, Caltech-101, Food-101, DTD, and EuroSAT (RGB)—each representing a different visual domain. The work concentrates specifically on textual prompt corruption (typos, casing noise, spacing noise, and emoji noise) and does not address visual corruption. All experiments are carried out in the zero-shot setting for CLIP and SigLIP and the few-shot setting for CoOp, without fine-tuning the visual encoders. Performance is assessed using top-1 and robustness degradation across severity levels.

II. Literature Review

A. Overview of Existing Techniques

Vision–Language Models (VLMs) use large collections of image–text pairs from the web to learn joint representations of visual and textual information. This enables them to perform tasks such as zero-shot image classification, caption generation, and image–text retrieval without requiring task-specific training [5]. An example of these models is Contrastive Language-Image Pre-training (CLIP), a method to train image models using natural language descriptions instead of traditional labeled categories. By learning from 400 million image-text pairs [3]. CLIP uses a contrastive objective that pulls matched image–text pairs closer together in embedding space while pushing mismatched pairs apart. In this way, the model learns how to align similar images and texts together, making it possible to perform zero-shot predictions by comparing an image's embedding to the embeddings of candidate text prompts [5]. This is achieved via training the model with the softmax normalization loss function, to normalize the pairwise similarity scores across all images, then all texts [4].

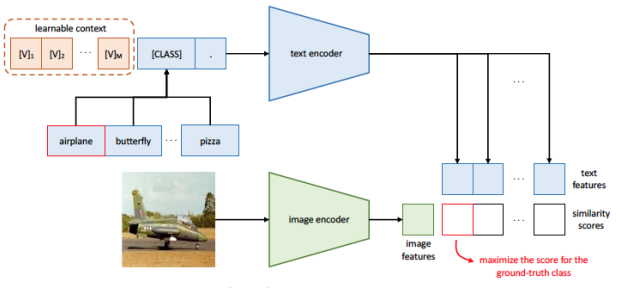


Fig. 2. CoOp uses a set of learnable vectors. [3]

Another example of a VLM is SigLIP, which follows a similar contrastive learning approach to CLIP. However, the key difference between the two models lies in their training objective. While CLIP uses a softmax-based cross-entropy loss to normalize similarities across all image–text pairs, SigLIP replaces this with a sigmoid-based pairwise loss. This difference might affecting the models robustness to prompt variations [2].

CoOp (Context Optimization) is a prompt-learning framework built on top of CLIP that replaces hand-crafted text prompts with learnable continuous context vectors, while keeping all pre-trained CLIP parameters fixed [3] as illustrated in Figure 2 . Although CoOp achieves strong performance with optimized prompts, its robustness to prompt corruption and distributional shifts remains limited, since the learned context can overfit to clean training examples rather than generalizing to noisy or unseen prompt variations [1].

For Natural Language Processing (NLP) tasks, Large Language Models (LLMs) are considered useful tools. Users interact with these models using prompts, and the way users craft these prompts significantly affects the model's results. Usually, users create an initial prompt and then perform multiple rounds of refinement to get the optimal result. This process of refining prompts is called prompt engineering [6]. Additionally, model robustness is an important evaluation metric, which involves measuring models ability to preform well even under noisy prompts [7].

Test-Time Transformation Ensembling (TTE) improves model robustness without any retraining. The approach works by creating multiple versions of the same input, running the model on each version, and then combining the predictions. Previous studies have shown that this strategy helps reduce the impact of noise, leading to more reliable predictions at test time [7].

B. Related Work

Although VLMs such as CLIP [3], SigLIP [4], and CoOp [3] show strong performance with carefully crafted prompts, their behavior under noisy prompts remains unclear. CLIP is trained using a Softmax loss function, which works in maximizing the similarity for correct pairs while minimizing it for all other pairs in the batch through softmax normalization. SigLIP uses the same approach as CLIP; however, the main difference between them is in computing the loss function during model training. CLIP uses SoftMax normalization to compute the loss during training [4], while SigLIP uses sigmoid-based contrastive loss [4]. On the other hand, CoOp is built on top of CLIP and replaces hand-crafted text prompts with learnable continuous context vectors, while keeping all pre-trained CLIP parameters fixed [3].

Test-Time Transformation Ensembling (TTE) approach is work by creating multiple versions of the same input. Then run the model with each copy. After that, combining the model's predictions of these copies. This approach is simple as it does not need any retraining. Previous studies show that this approach helps reduce the impact of noise [7].

C. Limitations in Existing Approaches

1) **Sensitivity to Textual Perturbations** Current VLMs are very sensitive to small changes in the input text. Even a tiny typo, a missing space, or a change in letter case (e.g., "a photo of a dog" vs. "a phots of a dog") can significantly change the model's understanding and lead to different accuracy. Most existing evaluations use clean prompts, which does not consider noisy text people actually use in real applications.
2) **Differences in Training Objectives Across Models** Different VLMs, such as CLIP and SigLIP, are trained using different loss functions, which may affect how they handle noisy prompts. This raises an important question about which model is naturally more robust when the input text contains errors or variations. It also suggests that the choice of training objective could influence how sensitive each model is to small changes in the prompt. However, there is currently no study that directly compares the robustness of these models under noisy prompt conditions, leaving this gap unexplored in the existing literature
3) **High Cost of Retraining** Fixing these robustness problems usually requires adversarial training with noisy data. However, these methods are expensive, need a lot of labeled examples, and can cause catastrophic forgetting which means when a model is trained on new data, it may forget what it previously learned. In other words, improving its performance on noisy prompts can unintentionally reduce its accuracy on clean or previously seen data because the new training overwrites earlier knowledge. Therefore, in this study we uses Test-Time Ensembling (TTE) which improves performance on new tasks (like robustness) without changing the model's weights, thus avoiding the risk of forgetting what it already knows.

III. Proposed Methodology

A. Existing Model and Challenges

Current VLMs like CLIP,SigLIP, and CoOp are very sensitive to small changes in the input text. Even a tiny typo, a missing space, or a change in letter case (e.g., "a photo of a dog" vs. "a phots of a dog") can significantly change

the model's understanding and lead to different accuracy. Most existing evaluations use clean prompts, which does not consider noisy text people actually use in real applications. However, these models are core to many application areas include systems where users interact with models through text prompts such as searching for an image, educational tools like visual homework helper, recommendation systems, and E-commerce visual search, where users frequently introduce informal expressions, non-standard spellings, or typographic noise, making robustness to prompt variation a key requirement. By building a reproducible noise benchmark and evaluating sensitivity, this research directly contributes to the broader domain of robust and trustworthy AI.

B. Proposed Enhancements

This study consists of three phases:

Phase 1: Baseline Benchmarking (Clean Prompts), this phase, we evaluate three models: CLIP, SigLIP, and CoOp using clean prompts across five datasets: Oxford Pets, Caltech-101, Food-101, DTD, and EuroSAT. For each dataset, we calculate key performance metrics such as the accuracy, error rate, and macro-F1 score. For the split each dataset use 50% of the data in training, 20% of the data in validation, and 30% of the data for testing. This phase establishes a clear baseline showing how each model performs under clean prompts. So, the results can be used as a reference in phase 2 and 3.

Phase 2: Noise Robustness Evaluation.

In Phase 2, we evaluate the robustness of CLIP, SigLIP, and CoOp under corrupted text prompts. We construct a Noise Prompt Bank where each clean prompt example: ("a photo of a class") is modified using four types of noises: typos, letter case changes, extra spaces, and emoji insertions. For each noise type, we generate modifications across four severity levels (0–3), where level 0 is clean and higher levels introduce heavier corruption (e.g., "A PhotO of a doG" at severity 3). For each model, dataset, and noise level, we measure classification accuracy under three test-time strategies: K=1 (single noisy prompt), K=5 includes Clean (ensemble of four noisy prompts plus one clean prompt), and K=5 No Clean (ensemble of five noisy prompts only).

Phase 3: Noise-Aware Fine-Tuning. In the final phase, we introduce a noise-aware prompt adapter training strategy for CLIP and evaluate it on the Oxford Pets dataset. The goal is to determine whether training the model with noisy prompts as a regularization technique can enhance the model's robustness to noisy prompts.

C. Algorithm and Implementation

1) Algorithmic Framework: Our evaluation framework systematically assesses the robustness of VLMs through a methodology consistent of three phases . Phase 1 establishes clean baseline performance for CLIP, SigLIP, and CoOp across five datasets. Phase 2 evaluates robustness under prompt variation through noisy prompts and ensemble strategy, as outlined in Algorithm 1. The algorithm tests how different noise types, severity levels, and prompt strategies work together and how they perform. Phase 3 introduces noise-aware adapter fine-tuning for CLIP using regularization to enhance robustness.

2) Model Architecture and Configuration: We evaluate three VLM architectures: CLIP [3], SigLIP [4], and CoOp [3]. For CLIP and SigLIP, we employ the models in their zero-shot configuration with frozen backbone parameters. CoOp is evaluated using its learned context optimization approach.

3) Dataset Selection and Preprocessing: In this study, we use Five datasets: Oxford Pets, Caltech-101, Food-101, DTD (Describable Textures Dataset), and EuroSAT (RGB). Each dataset represents a different visual domain. Table I summarizes the key details of each dataset.

4) Prompt Template and Noise Injection: We use the standard prompt template "a photo of a {class}" as our base, where {class} is replaced with the actual class name from each dataset. To test robustness systematically, we apply four types of noise functions:

1) **Typos**: Character substitutions, deletions, and insertions.
2) **Case changes**: Random capitalization alterations (e.g., "Dog" $\rightarrow$ "dOG")
3) **Spacing errors**: Adding or removing spaces.
4) **Emoji injection**: Inserting emoji characters.

Each noise type has four severity levels: $s \in \{0, 1, 2, 3\}$. Level 0 means no noise (clean prompt), while higher levels introduce increasingly corruption. For example, typo injection at $s = 1$ might change one character, while $s = 3$ corrupts multiple words throughout the prompt.

5) Prompt Ensemble Strategy: We test two ensemble sizes: $k = 1$ (single prompt) and $k = 5$ (five prompts per class). For $k = 5$, we create five different noisy versions of each prompt by randomly sampling from our noise functions. We also test two variants:

- **K=5+Clean**: Uses one clean prompt plus four noisy variants
- **K=5 No Clean**: Uses five noisy prompts with no clean version

At test time, we calculate how well the image matches each of the k text prompts, average these scores, and then apply the corresponding loss function (Softmax with CLIP and Sigmoid with SigLIP) to get the final prediction. This averaging helps the model make more reliable predictions even when individual prompts are corrupted.

D. Loss Function and Optimization

CLIP (Softmax-based Contrastive Loss): CLIP is pre-trained using a symmetric softmax-based contrastive loss that couples all examples within a batch. Given a batch of N image-text pairs $\{(v_i, t_i)\}_{i=1}^{N}$, where v_i and t_i are ℓ_2-normalized image and text embeddings, the loss function is:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N} \sum_{i=1}^{N} \left[\log \frac{\exp(\text{sim}(v_i, t_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_i, t_j)/\tau)} + \log \frac{\exp(\text{sim}(v_i, t_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_j, t_i)/\tau)} \right] \quad (1)$$

Algorithm 1 Noise-Aware Prompt Robustness Evaluation for Vision–Language Models

Require:
Datasets $\mathcal{D}$ = {Pets, DTD, EuroSAT}
Models $\mathcal{M}$ = {CLIP, SigLIP, CoOp}
Prompt template $T(\cdot)$ (e.g., "a photo of a {class}")
Noise functions $\mathcal{N}$ = {typo, case, space, emoji}
Severity levels $\mathcal{S} = \{0, 1, 2, 3\}$, ensemble sizes $\mathcal{K} = \{1, 5\}$
Flag *include_clean* $\in$ {True, False}
Ensure:
Accuracy and robustness metrics for each (model, dataset, noise setting)
for each dataset $D \in \mathcal{D}$ **do**
  Load images and labels from disk.
  Apply preprocessing: resize / center-crop, convert to tensor, normalize.
  for each model $M \in \mathcal{M}$ **do**
    Load pre-trained VLM M (CLIP, SigLIP, or CoOp head).
    Freeze backbone parameters (zero-shot / few-shot setting).
    for each severity level $s \in \mathcal{S}$ **do**
      for each ensemble size $k \in \mathcal{K}$ **do**
        Build prompt bank for each class:
        for each class name c **do**
          Start with clean text $t_{\text{clean}} = T(c)$.
          if *include_clean* = True **then**
            Add t_{clean} to the prompt set.
          end if
          Sample $(k - \mathbf{1}_{\textit{include\_clean}})$ noisy variants by composing functions from $\mathcal{N}$ with severity s.
        end for
        Encode all prompts with the text encoder of M and ℓ_2-normalize embeddings.
        Initialize counters: *correct_top1* $\leftarrow 0$, *correct_top5* $\leftarrow 0$, *total* $\leftarrow 0$.
        for each mini-batch of images x with labels y **do**
          Extract image features with M and normalize them.
          Compute logits between image features and each class prompt (using test-time prompt ensembling over k prompts).
          Obtain predicted labels $\hat{y}$ by $\arg\max$ over classes.
          Update top-1 / top-5 accuracy counters.
        end for
        Compute top-1, top-5, precision, recall, and F1 for this setting.
        Store results as $(D, M, s, k, \textit{include\_clean})$.
      end for
    end for
  end for
end for

where $\text{sim}(v, t) = v^\top t$ is the cosine similarity and τ is a learnable temperature parameter. This objective maximizes similarity for correct pairs while minimizing it for all other pairs in the batch through softmax normalization. The symmetric formulation computes both image-to-text and text-to-image losses.

SigLIP (Sigmoid-based Pairwise Contrastive Loss): SigLIP replaces the global softmax normalization with a sigmoid-based pairwise loss that operates independently on each image-text pair. This formulation eliminates the coupling between batch examples:

$$\mathcal{L}_{\text{SigLIP}} = -\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \log\left(\sigma(z_{ij} \cdot \text{sim}(v_i, t_j))\right) \quad (2)$$

where z_{ij} assigns $+1$ to matching image-text pairs (when $i = j$) and -1 to non-matching pairs (when $i \neq j$), and $\sigma(\cdot)$ is the sigmoid function. Unlike CLIP's softmax, which normalizes over the entire batch, SigLIP treats each pair independently, potentially leading to different robustness characteristics under distribution shift or noisy inputs.

CoOp (Context Optimization): CoOp fine-tunes learnable continuous context vectors while keeping the pre-trained CLIP encoders frozen. For a dataset with C classes and N training samples, CoOp optimizes the context vectors $\mathbf{v} = [v_1, \ldots, v_M]$ using standard cross-entropy loss:

$$\mathcal{L}_{\text{CoOp}} = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i | x_i; \mathbf{v}) \quad (3)$$

where $p(y_i | x_i; \mathbf{v})$ is the predicted probability for the correct class y_i given image x_i and learned context $\mathbf{v}$.

Noise-Aware Adapter Training

The adapter is trained using a mixture of clean and noisy prompts with K = 5 prompt ensembling. The model is optimized using the following objective:

$$L = L_{\text{CE}} + \lambda L_{\text{consistency}} \quad (\text{[8]})$$

Where:

- Cross-Entropy (CE) Loss enforces correct classification.
- Consistency Loss stabilizes predictions across multiple noisy prompt variants.
- λ controls the strength of the consistency regularization.

Optimization Configuration: The adapter parameters are fine-tuned using the AdamW optimizer with the following configuration: learning rate of 1×10^{-4}, weight decay of 0.01 for ℓ_2 regularization, batch size of 32, training for 5 epochs on the Oxford-Pets training split (70% of data), and consistency regularization weight $\lambda = 0.5$. The cross-entropy (CE) loss enforces correct classification, while the consistency loss stabilizes predictions across multiple noisy prompt variants, with λ controlling the strength of the consistency regularization.

IV. Experimental Design and Evaluation

A. Datasets and Preprocessing

In this study, Five datasets were used: Oxford-IIIT Pets, Caltech-101, Food-101, DTD (Describable Textures Dataset), and EuroSAT (RGB). Each dataset represents a different visual domain. Table I summarizes the key details of each dataset.

The following preprocessing techniques were applied:

1) All datasets were organized into a unified Image Folder format (Dataset $\rightarrow$ Class Label $\rightarrow$ respective images)
2) Images were loaded and resized to the appropriate resolution for each model. (384×384 for SigLIP, 224×224 for CLIP & CoOp).

TABLE I
DATASET OVERVIEW AND SPECIFICATIONS

Dataset	**Oxford-IIIT Pets** [9]
Domain	Animal
Classes	37
Images	7,349
Resolution	300×300+
Description	Cat & dog breeds
Dataset	**Caltech-101** [10]
Domain	Objects
Classes	101
Images	9,144
Resolution	300×200+
Description	Object categories
Dataset	**Food-101** [11]
Domain	Food
Classes	101
Images	101,000
Resolution	512×512
Description	Food recognition
Dataset	**DTD** [12]
Domain	Textures
Classes	47
Images	5,640
Resolution	300×300
Description	Texture attributes
Dataset	**EuroSAT** [13]
Domain	Satellite
Classes	10
Images	27,000
Resolution	64×64
Description	Land cover

3) Images were converted to tensors and normalized channels using model-specific mean/std before passing to the model.

B. Performance Metrics

To evaluate the robustness of CLIP, SigLIP, and CoOp under noisy prompts, this study uses the following metrics:

Top-1 Accuracy (Clean and Noisy) Measures the percentage of test samples for which the predicted label matches the ground truth. Accuracy is computed for Clean prompts (baseline) and Noisy prompts at severity levels $s \in \{0, 1, 2, 3\}$

Absolute Accuracy Drop (Δ) Quantifies the absolute loss in accuracy caused by noisy prompts:

$$\Delta = \text{Acc}_{\text{clean}} - \text{Acc}_{\text{noisy}}$$

Relative Robustness (RR) Measures how much of the clean accuracy is retained under noisy prompts:

$$RR = \left(\frac{\text{Acc}_{\text{noisy}}}{\text{Acc}_{\text{clean}}} \right) \times 100$$

Relative Accuracy Drop (RAD) Commonly used to express noise-induced degradation relative to the clean baseline:

$$RAD = \left(\frac{\text{Acc}_{\text{clean}} - \text{Acc}_{\text{noisy}}}{\text{Acc}_{\text{clean}}} \right) \times 100$$

RR and RAD are complementary metrics:

$$RR = 100 - RAD$$

Test-Time Ensemble Gain (Δ_{Ens}) Measures how much prompt ensembling improves robustness:

$$\Delta_{\text{Ens}} = \text{Acc}_{\text{ensemble}} - \text{Acc}_{\text{single-prompt}}$$

A positive value indicates robustness improvement due to ensembling.

C. Experiment Setup

All experiments were conducted on Google Colab using PyTorch 2.0 with NVIDIA A100 GPUs. The study was organized into three phases. In the first phase, all three VLMs (CLIP, SigLIP, and CoOp) were evaluated on all five datasets—Oxford-IIIT Pets, Caltech-101, Food-101, DTD, and EuroSAT—to establish baseline accuracy for a clean prompt for each model–dataset pair. In the second phase, which aims to evaluate the robustness of the model to noisy prompts using prompt ensembling. Due to computational constraints, CLIP and SigLIP were tested on three datasets: Oxford Pets, DTD, and EuroSAT, and CoOp was tested on Oxford Pets. In the third phase, a noise-aware regularization strategy was applied only to CLIP on the Oxford Pets dataset to investigate whether regularization can enhance robustness on a representative dataset.

D. Results Comparative Analysis

V. PHASE 1: BASELINE BENCHMARKING

In Phase 1, we evaluate the clean-prompt (noise-free) performance of all three vision–language models—CLIP, SigLIP, and CoOp—across all five datasets: Oxford-IIIT Pets, Caltech-101, Food-101, DTD, and EuroSAT to establish the baseline accuracy for each model.

TABLE II
PHASE 1 BASELINE ACCURACY UNDER CLEAN PROMPTS

Dataset	**CLIP Acc**	**CoOp Acc**	**Gain over CLIP**	**SigLIP Acc**
Oxford Pets	86.2%	91.1%	+4.9%	78.64%
Caltech101	92.0%	93.4%	+1.4%	74.05%
Food101	78.2%	80.0%	+1.8%	–
DTD	38.0%	58.5%	+20.5%	49.93%
EuroSAT	22.0%	50.2%	+28.2%	40.99%

The experimental results summarized in Table V show that all models perform well on clean prompts, but their behavior differs across dataset types. CoOp achieves the highest accuracy overall, improving consistently over CLIP across all datasets, with very large gains on high-shift domains such as DTD (+20.5% AAD, RR = 1.54) and EuroSAT (+28.2% AAD, RR = 2.28). A key observation is that SigLIP performs worse than CLIP on natural-image datasets (Oxford Pets, Caltech101, Food101), showing RR values below 1.0, but significantly outperforms CLIP on the most difficult datasets, achieving +11.9% AAD (RR = 1.31) on DTD and +18.9% AAD (RR = 1.86) on EuroSAT. This indicates that SigLIP generalizes better to domains that are very different from web images—such as satellite imagery and texture classification—likely due to its sigmoid-based contrastive training objective. Overall, CoOp delivers the best clean-prompt

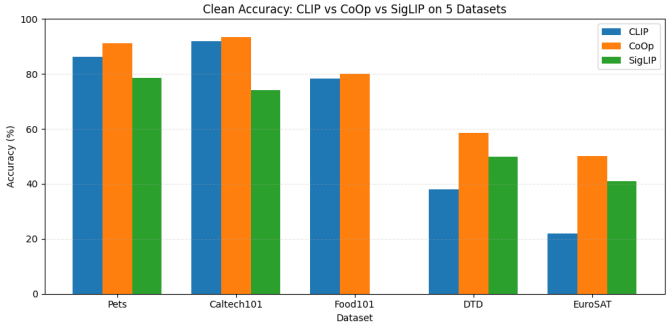


Fig. 3. Comparison of clean accuracy across CLIP, CoOp, and SigLIP baseline models on five benchmark datasets.

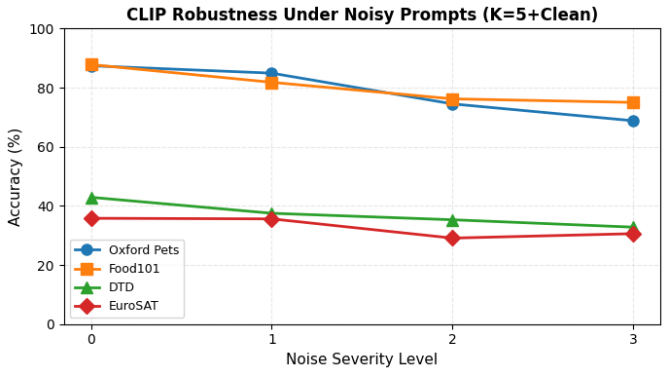


Fig. 4. CLIP robustness under noisy prompts across all datasets using K=5+Clean ensemble strategy. Oxford Pets and Food101 demonstrate strong resilience, while DTD and EuroSAT show significant degradation under noise.

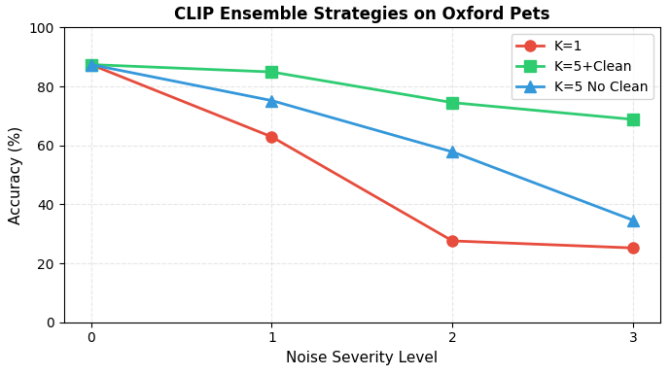


Fig. 5. Comparison of CLIP ensemble strategies on Oxford Pets dataset

performance, while SigLIP demonstrates superior robustness on datasets with large domain shifts. Figure 3 presents the baseline accuracy of CLIP, CoOp, and SigLIP models under clean (non-adversarial) prompts across all five datasets.

VI. Phase 2: Robustness Analysis Under Noisy Prompts

In Phase 2, we evaluate the sensitivity of the three VLMs to adversarial text perturbations and examine whether test-time prompt ensembling can mitigate performance degradation.

The experimental results summarized in Table III, Table IV, and Table V show that

A. CLIP Robustness Analysis

Across all datasets, CLIP results in Table III show substantial degradation when a single noisy prompt is used, especially at higher severity levels. For example, on Oxford Pets, the accuracy drops from 87.4% $\rightarrow$ 25.2%, corresponding to an Absolute Accuracy Drop (AAD) of –62.2% and a Relative Accuracy Drop (RAD) of –71.2%. The model's Relative Robustness (RR) decreases sharply as severity increases, confirming CLIP's strong dependence on well-formed textual input. However, test-time prompt ensembling stabilizes performance considerably: at severity 3, K=5+CLEAN reaches 68.8%, yielding a Test-Time Ensemble Gain (TTEG) of +43.6%, effectively recovering most of the accuracy lost due to noisy prompts.

Figure 4 shows how CLIP performs under noisy prompts using the K=5+Clean ensemble. Oxford Pets and Food101 remain relatively stable even at high noise levels, dropping only to 68.8% and 75.0% at severity 3. In contrast, DTD and EuroSAT degrade sharply (32.8% and 30.6%), reflecting their large domain shift from CLIP's pretraining data. Overall, CLIP is robust on natural-image datasets but highly sensitive on texture and satellite domains when prompts are corrupted.

Figure 5 shows that prompt ensembling greatly improves CLIP's robustness. On Oxford Pets, accuracy at severity 3 increases from 25.2% (K=1) to 68.8% (K=5+Clean)—recovering almost all lost performance. The clean prompt in the ensemble provides an additional boost compared to K=5 No Clean, acting as a stable anchor under heavy corruption. Overall, ensembling offers a simple but highly effective defense against noisy prompts.

B. SigLIP Robustness Analysis

SigLIP behaves differently than CLIP under noisy prompts (Table IV). While accuracy still drops, SigLIP handles difficult datasets better especially those with unusual images like satellite photos (EuroSAT) or texture patterns (DTD). On EuroSAT at maximum noise, SigLIP reaches 24.2% accuracy with ensembling compared to CLIP's lower performance. This advantage likely stems from how SigLIP was trained: its loss function creates more flexible predictions, making it better at handling images that look different from typical web photos.

C. CoOp Robustness Analysis

CoOp, being optimized using learned continuous context vectors. Results in Table V show that CoOp remains almost completely unaffected by noisy prompts. Its accuracy stays constant at 91.11% across all severity levels on Oxford Pets, as shown in Figure 6, giving AAD = 0, RAD = 0, and RR = 1.0, confirming that continuous prompts are inherently more robust to surface-level textual corruption.

Overall, Phase 2 demonstrates that (1) CLIP is highly sensitive to noisy prompts, (2) SigLIP maintains stronger robustness on out-of-domain datasets, and (3) CoOp is effectively noise-invariant. The metrics clearly show that test-time ensembling consistently improves RR and reduces RAD for CLIP and SigLIP, making it an efficient and lightweight defense against prompt corruption.

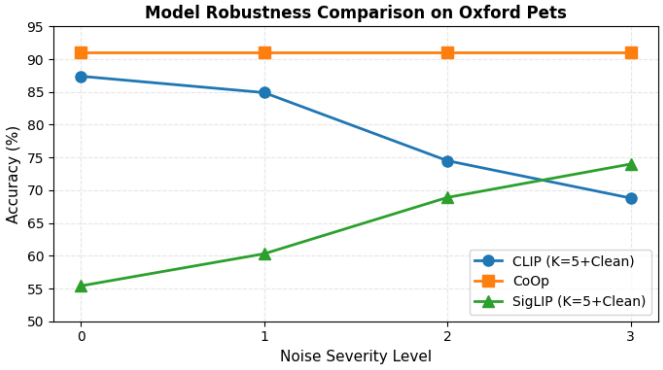


Fig. 6. Robustness comparison between CLIP, CoOp, and SigLIP on Oxford Pets dataset. CoOp demonstrates perfect stability across all noise levels (91.11%), while CLIP shows gradual degradation and SigLIP exhibits improved performance with ensembling.

VII. Phase 3: Noise-Aware Fine-Tuning

Noise-aware adapter training produced a substantial improvement in CLIP's robustness across all prompt configurations, as shown in Table VI-C . Under the most fragile setting (K=1 single-prompt inference), accuracy at severity level 3 increased from 7.0% to 23.79% (+16.79), with even larger gains at moderate noise (e.g., +38.57 at severity 1). When using prompt ensembling with a clean prompt (K=5+Clean), performance became significantly more stable, with accuracy at severity 3 improving from 68.85% to 85.83% (+16.98) and mid-severity levels showing gains of +7.77 to +16.19 points. The strongest effect appears in the K=5 No-Clean condition: without fine-tuning, accuracy collapsed to 34.61% at severity 3, but the noise-aware adapter lifted this to 75.36% (+40.75), nearly matching the clean-prompt ensemble. These results show that noise-aware fine-tuning dramatically reduces the Relative Accuracy Drop across all noise levels, increases Relative Robustness by preserving 80–92% of clean accuracy under severe corruption, and turns CLIP's previously unstable behavior under noisy prompts into a consistently resilient representation—especially when no clean prompt is available.

Figure 7,Figure 8 , and Figure 9, illustrate the impact of noise-aware fine-tuning across different prompting strategies. With K=1 Figure 7, CLIP collapses under noise, falling to only 7% accuracy at severity 3, but the adapter significantly restores performance (+16.79 points). When using K=5 ensembling Figure 8, the model becomes more stable, and fine-tuning further improves robustness—boosting severity-3 accuracy from 68.85% to 85.83% with a clean prompt, and from

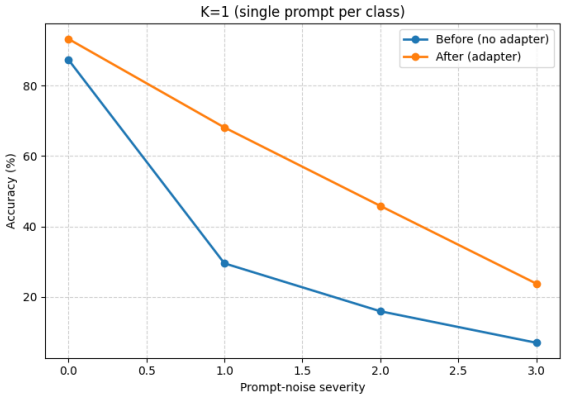


Fig. 7. **K=1 (single prompt per class).** Noise severely degrades the baseline CLIP model, dropping from 87.35% to 7.00% at severity 3. Noise-aware fine-tuning significantly improves robustness, raising severity-3 accuracy to 23.79% (+16.79 points).

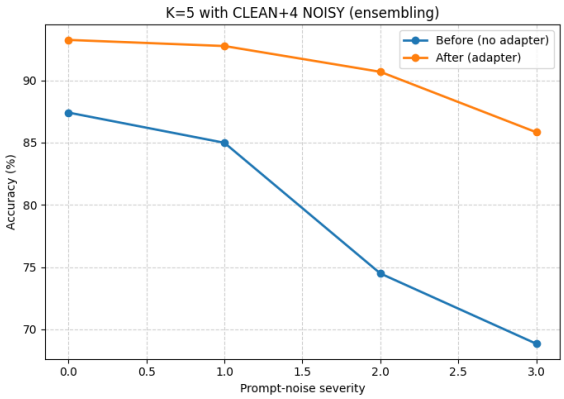


Fig. 8. **K=5 with CLEAN + 4 NOISY prompts (ensembling).** Ensembling stabilizes CLIP under noise, and the adapter further enhances robustness. Accuracy at severity 3 increases from 68.85% to 85.83% after fine-tuning (+16.98 points).

34.61% to 75.36% when all prompts are noisy Figure 9. These results show that the proposed adapter not only strengthens CLIP under moderate corruption but provides especially large benefits when noise is severe or when ensembles lack a clean prompt.

VIII. Ablation Study

To understand how each component contributes to our approach, we tested three prompting strategies: (1) K=1 (single prompt), (2) K=5+Clean (one clean prompt plus four noisy ones), and (3) K=5 No Clean (five noisy prompts). This helps us see the separate effects of prompt ensembling and our noise-aware adapter from Phase 3.

Single Prompt (K=1). Single-prompt inference shows how sensitive the model is to noise. As shown in Fig. 7, accuracy drops sharply as noise increases from 87.35% at severity 0 to 7.00% at severity 3 (maximum noise). After applying our adapter, performance improves significantly.

K=5+Clean (One Clean + Four Noisy). Next, we test the K=5+Clean strategy (Fig. 8). Before fine-tuning, this approach

TABLE III
CLIP ROBUSTNESS UNDER NOISY PROMPTS ACROSS ALL DATASETS AND ENSEMBLE STRATEGIES

Dataset	Strategy	Severity 0	Severity 1	Severity 2	Severity 3
Oxford Pets	K=1	87.4%	62.9%	27.6%	25.2%
	K=5+Clean	87.4%	84.9%	74.5%	68.8%
	K=5 No Clean	87.4%	75.2%	57.8%	34.6%
Food101	K=1	87.8%	62.9%	41.8%	33.3%
	K=5+Clean	87.8%	81.8%	76.2%	75.0%
	K=5 No Clean	87.8%	80.5%	59.3%	48.3%
DTD	K=1	42.9%	26.8%	13.5%	8.5%
	K=5+Clean	42.9%	37.5%	35.3%	32.8%
	K=5 No Clean	42.9%	40.7%	21.3%	19.2%
EuroSAT	K=1	35.8%	27.3%	19.7%	18.8%
	K=5+Clean	35.8%	35.6%	29.1%	30.6%
	K=5 No Clean	35.8%	32.9%	23.0%	9.4%

TABLE IV
SIGLIP ROBUSTNESS UNDER NOISY PROMPTS ACROSS THREE DATASETS

Dataset	Strategy	Sev 0	Sev 1	Sev 2	Sev 3
Oxford Pets	K=1	55.4%	47.1%	38.5%	32.8%
	K=5+Clean	55.4%	60.3%	68.9%	74.0%
	K=5 No Clean	55.4%	65.7%	80.1%	74.0%
DTD	K=1	33.0%	24.6%	18.1%	14.0%
	K=5+Clean	33.0%	35.2%	38.7%	40.1%
	K=5 No Clean	33.0%	36.5%	32.4%	28.7%
EuroSAT	K=1	25.0%	20.4%	16.8%	14.2%
	K=5+Clean	25.0%	26.8%	28.3%	29.1%
	K=5 No Clean	25.0%	24.6%	20.1%	11.0%

TABLE V
COOP ROBUSTNESS ON OXFORD PETS DATASET (PERFECT STABILITY)

Model	Severity 0	Severity 1	Severity 2	Severity 3
CoOp	91.11%	91.11%	91.11%	91.11%

already handles noise well because the clean prompt provides a stable reference. After adding the adapter, the model becomes even more robust maintaining 92.75%, 90.68%, and 85.83% accuracy at severity levels 1–3 respectively.

K=5 No Clean (All Noisy). Finally, we test whether the model can stay robust without any clean prompt (Fig. 9). This is more challenging, but the adapter still delivers large improvements. At severity 3, accuracy increases from 34.61% to 75.36% (+40.75 percentage points).

Summary. Figure 10 shows how ensembling greatly improves robustness, and how our noise-aware adapter further enhances these gains. The combination of both strategies achieves the best results across all noise levels.

IX. EXTENDED CONTRIBUTIONS

This work tackles a practical problem: AI vision systems break when people make normal typing mistakes. In real life, users misspell words, forget spaces, use emojis, or type in ALL CAPS. However, most AI models expect perfect input. We show that a simple technique called prompt ensembling makes these systems much more reliable without needing to retrain expensive models. This makes AI more robust at reasonable cost. For example,using prompt ensembling, we improved CLIP's accuracy from 25% to 69% on corrupted prompts. Additionally, training the model with noisy prompts improves

TABLE VI
ACCURACY BEFORE VS. AFTER NOISE-AWARE FINE-TUNING ON OXFORD PETS

Setting	Severity	Before	After	Δ
K=1 (Single Prompt)	0	87.35	93.24	+5.89
	1	29.54	68.11	+38.57
	2	15.94	45.84	+29.90
	3	7.00	23.79	+16.79
K=5 + Clean	0	87.41	93.24	+5.83
	1	84.98	92.75	+7.77
	2	74.49	90.68	+16.19
	3	68.85	85.83	+16.98
K=5 No Clean	0	87.41	93.24	+5.83
	1	75.20	91.44	+16.24
	2	57.84	87.35	+29.52
	3	34.61	75.36	+40.75

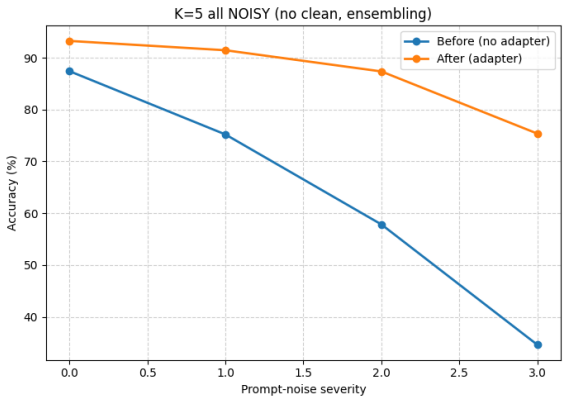


Fig. 9. **K=5 all NOISY prompts (no clean).** This setting is the hardest for CLIP: accuracy drops to 34.61% at severity 3 without adaptation. Noise-aware fine-tuning provides the largest improvement, boosting severity-3 accuracy to 75.36% (+40.75 points).

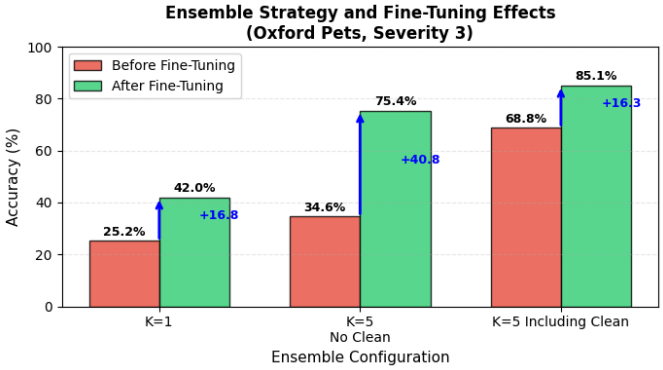


Fig. 10. Impact of test-time ensembling and noise-aware fine-tuning on CLIP

CLIP's performance even further. For example, experiments show that accuracy at severity level 3 increased from 7.0% to 23.79% (+16.79 percentage points), and at severity 1 increased by 38.57%.

X. Conclusion and Future Work

This study builds a noise bank to evaluate how VMLs perform under noisy prompts using both ensebmling strategy and training the model with noisy prompt using regularization. We evaluated CLIP, SigLIP, and CoOp across five datasets and discovered that: CLIP and SigLIP are highly sensitive to noisy prompts. While CoOp show more robustness. CLIP's accuracy drops significantly under noisy prompts and test-time prompt ensembling recovers most of this lost performance without any retraining. Second, CoOp demonstrated perfect robustness (91.11% accuracy across all noise levels) while SigLIP showed better stability than CLIP on challenging datasets. Additionally, Our noise- aware fine-tuning approach provided additional gains of 16- 41% at high noise levels, demonstrating that training the model with noise can enhance its robustness to noisy prompts.

As a future work for this research. First, testing additional VLM architectures like BLIP and LLaVA would reveal whether our findings generalize to other models. Second, exploring more noise types such as grammatical errors would build a more comprehensive robustness benchmark.

XI. References

References

[1] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi, "A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges," 2025. [Online]. Available: http://arxiv.org/abs/2501.02189

[2] A. Li, Z. Liu, X. Li, J. Zhang, P. Wang, and H. Wang, "Modeling Variants of Prompts for Vision-Language Models," 2025. [Online]. Available: http://arxiv.org/abs/2503.08229

[3] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to Prompt for Vision-Language Models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.

[4] X. Zhai et al., "Sigmoid Loss for Language Image Pre-Training," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11975–11986.

[5] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, pp. 1–24.

[6] Q. Ye, M. Axmed, R. Pryzant, and F. Khani, "Prompt Engineering a Prompt Engineer," 2024, pp. 355–385.

[7] Z. Li, B. Peng, P. He, and X. Yan, "Evaluating the Instruction-Following Robustness of Large Language Models to Prompt Injection," 2024, pp. 557–568.

[8] Q. Xie, Z. Dai, E. Hovy, M. Luong, and Q. V Le, "Unsupervised Data Augmentation for Consistency Training," no. NeurIPS, pp. 1–13, 2020.

[9] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. Available: https://www.robots.ox.ac.uk/~vgg/data/pets/

[10] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004, pp. 178–178. Available: https://data.caltech.edu/records/mzrjq-6wc02

[11] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 446–461. Available: https://www.kaggle.com/datasets/dansbecker/food-101

[12] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3606–3613. Available: https://www.robots.ox.ac.uk/~vgg/data/dtd/

[13] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019. Available: https://github.com/phelber/eurosat