Dual Polarity Prompts with Stochastic Entropy Perturbation for Label Noise

Changhui Hu[1,†]
chhuhuxx121@alumnes.ub.edu
Bhalaji Nagarajan[2,†]
bhalaji.nagarajan@bsc.es
Ricardo Marques[3,‡]
ricardo.marques@upf.edu
Petia Radeva[1,4,‡]
petia.ivanova@ub.edu

[1] Dept. de Matemàtiques i Informàtica, Universitat de Barcelona
[2] Barcelona Supercomputing Center
[3] Grup de Recerca de Tecnologies Interactives (GTI), Universitat Pompeu Fabra (UPF)
[4] Institut de Neurosciències, Universitat de Barcelona

Abstract

Pre-trained Vision-Language Models (VLMs) have been widely adopted across various applications through few-shot prompt tuning. However, their potential to effectively address label noise, a common challenge in real-world data, remains largely unexplored. While recent prompt tuning methods exhibit robustness under low to moderate noise, they struggle in handling higher noise ratios or more complex noise patterns. To bridge this gap, we introduce **"DP-SEP"**, a novel framework designed to enhance the adaptation of VLMs under challenging label noise. DP-SEP consists of two key components: (1) **Dual Polarity (DP)** Prompts, which dynamically adjust text-image alignments by bringing the *correct* description closer, while pushing away *incorrect* ones. DP includes a novel class reordering function to generate negative prompts and a Polarity Alignment Reweighting (PAR) factor to effectively balance the dual prompts. (2) **Stochastic Entropy Perturbation (SEP)**, which injects controlled randomness to reduce the VLM's overconfidence in noisy labels. Extensive experiments on 7 synthetic datasets and two real-world noisy benchmarks demonstrate that DP-SEP significantly outperforms existing methods. In particular, we achieve an impressive average gain of 11.17% and 16.98% under 65% and 80% sym. noise. Significant improvements of DP-SEP establishes as a new direction for handling label noise in adopting VLMs.

1 Introduction

Rapid advancement of large-scale model architectures [24, 45] and web-scale data [12, 33] has led to the development of powerful general-purpose models [29, 31] capable of excelling across diverse downstream tasks. In particular, large-scale Vision-Language Models (VLMs) such as CLIP (Contrastive Language-Image Pretraining) [28] and its variants [5, 35, 36, 37] leverage contrastive learning to align vision and language modalities within a shared embedding space, enabling a richer semantic understanding. These models have demonstrated

† Equal Contribution. ‡ Equal Supervision.

exceptional performance in image classification tasks, achieving strong zero-shot generalization [21, 27, 50]. However, effectively adapting CLIP to specific downstream tasks remains a significant challenge [1]. **Prompt Learning** (or **Prompt Tuning**) [18, 43] is an efficient alternative to fine-tuning models like CLIP. Instead of relying on manually crafted text prompts, prompt tuning optimizes a set of few *learnable* differentiable context vectors to better adapt models to specific tasks, using task-specific objectives such as cross-entropy for classification [49]. Prompt tuning helps to refine the understanding of models preserving its generalization capabilities [14], thus enabling effective task adaptation without the computational cost of full model fine-tuning.

Optimizing prompts, especially in VLMs such as CLIP, without overfitting to task-specific data remains an open challenge [20]. Several methods like CoCoOp [48], MaPLe [10] and PromptSRC [11], are proposed to enhance prompt optimization and improve adaptability. However, a critical and yet less-explored challenge in prompt tuning is its *susceptibility to label noise* in downstream datasets, which can significantly degrade performance and hinder adaptation. PTNL [41] show that prompt tuning can exhibit extraordinary robustness to noisy labels, largely due to the constraints imposed by language priors on visual representations and the inherent properties of the prompt mechanism itself. Despite its contributions, PTNL primarily focuses on stabilizing prompt tuning, leaving key challenges unaddressed. This includes handling more complex label errors, mitigating the model's overconfidence in incorrect labels, and improving robustness under extreme noise conditions. Recently, DeFT [40] used static detectors along with positive and negative prompts in a two stage process of first detecting wrong labels and then learning with noisy labels.

In this work, we introduce **DP-SEP**, a novel prompt-tuning framework designed to improve both robustness and adaptability in noisy label scenarios. Our approach is built upon two key components: (1) *Dynamic* **Dual Polarity (DP)** prompts, which introduce opposing prompt representations to reinforce the learning of correct semantic alignments while mitigating the influence of incorrect labels. By introducing DP - positive and *uniquely generated* negative prompts - we strengthen the alignment with correct descriptions while simultaneously pushing away incorrect ones. DP further includes a novel class reordering function to create negative prompts and uses **Polarity Alignment Reweighting (PAR)** Factor to effectively balance the dual prompts. (2) **Stochastic Entropy Perturbation (SEP)**, which dynamically regulates model confidence to prevent memorizing and overfitting to wrong labels during downstream fine-tuning.

Our main contributions are summarized as follows:

1. We introduce a novel **Dual Polarity prompting** strategy along with prompt balancing functions, that enhances image-text alignment of VLMs by simultaneously maximizing similarity with *correct* descriptions and minimizing similarity with *incorrect* ones, leading to robust discrimination in high and complex noise environments.

2. We propose **Stochastic Entropy Perturbation** loss which mitigates overconfidence in noisy labels by introducing stochastic perturbations, effectively reducing the model's memorization of noisy samples.

3. We perform an **extensive validation** of our method on seven synthetic benchmarks with symmetric and instance-dependent noise and two real-world noise benchmarks, covering a wide range of noise distributions, demonstrating the superior noise-robustness of our prompt-tuning method.

2 Related Works

Recent advances in VLMs, exemplified by CLIP [28], ALIGN [8], CoCa [44] and DeepSeek-VL[19] — pre-trained on massive web-scale datasets [2, 33] of image-text pairs — enable learning of a broad range of concepts and relationships, and consequently, demonstrate strong zero-shot performance. However, significant efforts have been dedicated to effectively fine-tune these large models [16, 32, 39] for improved performance on specific downstream tasks. Prompt learning, particularly for models like CLIP, offers an efficient alternative.

Prompt Tuning for VLMs: CLIP [28] mainly use manually constructed text templates (like "a photo of a {class}"), but subsequent study [49] shows slight changes in prompts significantly affect downstream performances. CoOp [49] proposed automatically optimizing learnable vectors thereby replacing manually engineered templates and significantly improving the model's adaptation efficiency to downstream categories. CoCoOp [48] introduced a lightweight network to extract key information from image features and integrate it into the prompt. MaPLe [10] explored the simultaneous maintenance of hierarchical prompt structures in both language and vision branches. KAPT [9] proposed integrating external knowledge descriptions in prompt learning. ProMetaR [25] optimized prompts through meta-regularization strategy, emphasizing the simultaneous constraints at different levels. PromptSRC [11] optimized prompts for tasks-specific and task-agnostic representations. DePT [47] decoupled base specific knowledge from feature channels during prompt tuning.

Noisy Labels in Prompt Tuning: Prompt tuning for VLMs has shown excellent adaptability and inherent robustness to label noise. However, optimizing prompts in the presence of label noise, is less explored. PTNL [41] explored why prompt tuning tolerates mild to moderate label noise compared to fine-tuning or linear probing, however, does not address high-noise ratios or complex noise patterns. JoAPR [7] introduced joint adaptive thresholds and label correction mechanisms to enhance prompts under label noise. DeFT [40] employed positive and negative prompts as static detectors and adopted a two stage method of first detecting and then learning to train with noisy labels. In this work, we introduce DP-SEP, explicitly designed for high-noise, complex and diverse annotation error scenarios. In contrast to other methods, we introduce novel dual polarity prompts construction and prompt balancing functions, along with stochastic perturbation, and propose an end-to-end learnable prompt tuning framework directly optimizing on noisy labels to mitigate its impact and to significantly improve downstream classification performances.

3 Our Method

The proposed **DP-SEP** method is a prompt-tuning framework that efficiently adapts large pre-trained vision-language models, such as CLIP, to downstream tasks with noisy labels. DP-SEP (see Figure 1) features two novel components. First, to enhance the discriminative capabilities of text prompts, we introduce **Dual Polarity** prompts: *positive* and *negative* prompts. Positive prompts are constructed following general CoOp [49], whereas Negative prompts are constructed by systematically reordering the class labels ensuring positional disparity. Second, to address label noise in downstream tasks, we propose a new loss function combining **Stochastic Entropy Perturabation**, derived from image-positive similarities,

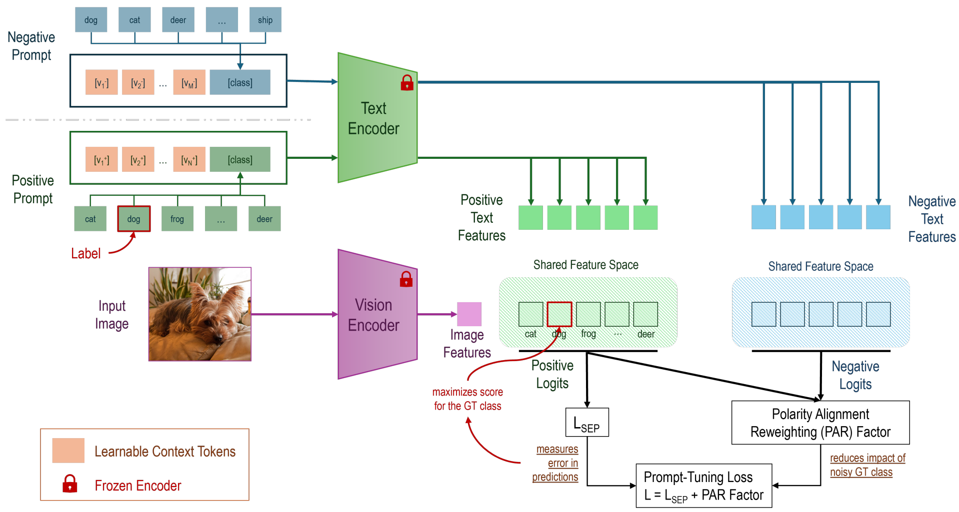


Figure 1: **Overview of DP-SEP:** *Dual polarity* prompts learn contrasting information and use various balancing functions to better focus the model adaptability in noisy label environments. *Stochastic perturbations* mitigate the memorization of label noise, commonly found in losses such as CE loss.

and **Polarity Alignment Reweighting Factor** leveraging both positive and negative similarities. The combined loss enables robust, noise-resistant prompt-tuning representations.

3.1 Dual Polarity Prompts

The core idea of DP prompts is that, in the presence of label noise, the similarity between an image and its *correct* class description will be significantly higher than that between the image and a *noisy* description within the learning space. We define two *polar* prompts, positive and negative, each learning the context vectors in a distinct manner. Both prompts are designed as unified context vectors [49], where the same context vector ctx is shared with all classes. We define **positive** prompt as a learnable context vector (ctx^{pos}), following CoOp [49], of size N and concatenate it with each class, c_i, for $i = 1, 2, \ldots, C$. We define our unique **negative** prompt by learning a separate context vector (ctx^{neg}) of size M, which, instead of concatenating with the same class, c_i, introduces a class reordering function $\pi(i)$ to ensure no class position is maintained as in the positive prompt. Thus, dual prompts yield:

$$P_i^{\text{pos}} = [ctx^{\text{pos}}] \oplus c_i, \qquad P_i^{\text{neg}} = [ctx^{\text{neg}}] \oplus c_{\pi(i)} \tag{1}$$

where $\oplus$ represents sequential concatenation and $\pi(i) \neq i$ for all i. For simplicity, we consider equal sized context vectors ctx^{pos} and ctx^{neg}, i.e., $M = N$. Given an image $\mathbf{x}$, we extract visual features $\mathbf{v} = V(\mathbf{x})$ from the vision encoder, V. We obtain text features $\mathbf{z}^{\text{pos}}$ and $\mathbf{z}^{\text{neg}}$ from the text encoder, T, for the positive prompts, P^{pos}, and negative prompts, P^{neg}, respectively. DP prompts are designed to increase the similarity between the image and its correct textual features, while decreasing the similarity between the image and noisy features. The key benefits of DP prompts are:

- *Enhancement of contrast signal:* Simultaneous existence of positive and negative prompts provides two opposing text embedding ($\mathbf{z}^{\text{pos}}$ vs. $\mathbf{z}^{\text{neg}}$) for an image x, creating effective contrast learning structure to pull the positive prompts closer and the negative prompts farther away.

- *Noise resistance:* Even when the label y_i is incorrect, randomly assigned negative prompts still tend to help to learn clear decision boundaries, reducing the dependence on individual labels.

- *Richness:* The dynamic DP extends beyond single images, single class scenarios, by randomly associating an incorrect class with each class. This ensures that the model is constantly exposed to contrast examples of "incorrect descriptions", reinforcing its ability to distinguish between incorrect semantic associations.

3.1.1 Polarity Alignment Reweighting Factor

To effectively balance the DP prompts in the feature space, we introduce *Polarity Alignment Reweighting* **(PAR)** factor. We define three alignment objectives, each focusing on the distance or distributional relationship between the image-positive prompt and image-negative prompt in the feature space. These objectives enforce specific alignment constraints by either narrowing or widening the similarity between different feature.
(1) Positive Alignment (A^{pos}) focuses on aligning image features v and positive prompt features z^{pos}, ensuring that the model effectively captures the shared semantics between the image and its correct category or accurate textual description. We define A^{pos} as:

$$A^{\text{pos}} = \frac{1}{N}\sum_{n=1}^{N} \max(0, \mu^{\text{pos}} - s^{\text{pos}}(x_n, c_n)), \tag{2}$$

$s^{\text{pos}}(x_n, c_n) = \text{softmax}(\alpha v_n \cdot z_n^{pos})$, μ^{pos}: threshold ensuring if $s^{\text{pos}}(x_n, c_n)$ falls below μ^{pos}, the model is encouraged to increase similarity between image and correct class embeddings.
(2) Negative Alignment (A^{neg}) focuses on ensuring the image features remain distinct from negative prompt features, preventing the model from mistakenly aligning images with incorrect class descriptions. A^{neg} provides a distinction between correct and incorrect pairs and measures the gap between positive and negative features. We define the A^{neg} as:

$$A^{\text{neg}} = \frac{1}{N}\sum_{n=1}^{N} \max(0, s^{\text{neg}}(x_n, c_{\pi(n)}) - s^{\text{pos}}(x_n, c_n) + \mu^{\text{neg}}), \tag{3}$$

$s^{\text{neg}}(x_n, c_{\pi(n)}) = \text{softmax}(\alpha v_n \cdot z_n^{\text{neg}})$, μ^{neg}: threshold (set 0.5 empirically) ensuring sufficient gap between positive and negative similarities forcing model to push negative descriptions further away.
(3) Modality Coupling (MC) introduces an additional coupling constraint between image and text features (both positive and negative prompts), ensuring that the model considers the overall structural differences between the prompts. We define the *MC* as:

$$M_C = \frac{1}{N}\sum_{n=1}^{N} \max(0, s^{\text{pos}}(x_n, c_n) - s^{\text{neg}}(x_n, c_{\pi(n)})). \tag{4}$$

MC ensures positive prompt consistently maintains higher similarity than negative prompt.

3.2 Stochastic Entropy Perturbation

When fine-tuning models like CLIP, shared embedding space of image and text features is adapted to minimize task-specific loss, such as cross-entropy (CE). The predicted probability, p, of a label, y, is derived from the similarity between image features, v and text features, z. This essentially shifts the learning objective from original contrastive learning of CLIP to a more discriminative one. One of the prominent problems of learning with noisy labels is the tendency of models to almost perfectly memorize all samples (including those with wrong labels) resulting in poor generalization [46]. To tackle this problem, we introduce a stochastic entropy perturbation term: $-\frac{1}{N}\sum_{n=1}^{N}(p_n+\delta_n)l_n$, designed to prevent the model from being too confident about certain class predictions. N is the number of data samples, $l_n = log(p_n)$ and $\delta_n \sim N(0,\sigma^2)$ is the actual stochastic perturbation introducing controlled noise to the entropy term during each forward pass, discouraging the model from converging to an overly sharp loss distribution. Finally, our proposed **SEP Loss** over the similarity between image-positive prompts is defined as follows:

$$\mathcal{L}_{SEP} = -\lambda\frac{1}{N}\sum_{n=1}^{N}(p_n+\delta_n)l_n - (1-\lambda)\frac{1}{N}\sum_{n=1}^{N}y_n l_n, \tag{5}$$

$-\frac{1}{N}\sum_{n=1}^{N}y_n l_n$: Standard CE that supervises labeled samples, λ controls the influence of SEP on the overall loss. By introducing stochastic entropy intervention, we penalize overly concentrated prediction distributions, preventing the model from treating noisy labels as *real learning signals*. By maintaining higher entropy for potentially mislabeled samples, we reduce the gap between the empirical and target risks. This encourages the model to be *less confident* on suspected noise and can result in a better approximation to the true distribution.

3.3 Overall Training

The overall DP-SEP loss, $\mathcal{L}_{\mathcal{DP}-\mathcal{SEP}}$, is formulated as:

$$\mathcal{L}_{\mathcal{DP}-\mathcal{SEP}} = \mathcal{L}_{SEP} + A^{\text{pos}} + \alpha A^{\text{neg}} + \beta M_C, \tag{6}$$

where α is the weight of A^{neg}, dynamically adjusted during training epochs $e = \{1,2,\ldots,E\}$ so that: $\alpha = \alpha_{\min} + (1-\alpha_{\min})(e/E)$. β is the weight of objective M_C.

4 Experiments

We perform extensive experiments across 7 synthetic benchmarks, ImageNet [30], Caltech101 [6], Oxford-IIIT Pets [26], Flowers102 [23], Food101 [3], DTD [4], UCF101 [34], alongside 2 large real-world benchmarks, WebVision [17] and Food101N [15]. We compare the performance of DP-SEP with PTNL [41], our baseline method. Comprehensive experimental settings are detailed in Supplement Section A.

4.1 Performance Comparison

Performance under Synthetic Noise: We present the performance comparisons under synthetic noise in Table 1. As a general tendency, models show a decline in performance as noise ratio increases. DP-SEP outperforms PTNL across a wide range of symmetric (sym.)

Dataset	Method	Symmetric 12.5%	25%	50%	65%	80%	Instance 20%	40%	60%	80%
ImageNet	PTNL	65.52	64.74	62.33	59.14	55.18	64.83	62.83	60.09	53.73
	DP-SEP	**65.63**	**65.38**	**64.59**	**63.93**	**63.02**	**65.30**	**64.58**	**63.99**	**62.95**
Caltech101	PTNL	94.29	92.84	91.22	90.13	83.29	92.99	91.42	88.95	82.63
	DP-SEP	**94.90**	**94.57**	**94.39**	**93.10**	**92.76**	**94.62**	**94.59**	**94.15**	**92.54**
OxfordPets	PTNL	83.60	81.86	74.70	56.93	50.43	81.60	79.20	70.05	29.27
	DP-SEP	**86.38**	**85.13**	**85.08**	**84.94**	**83.54**	**86.15**	**85.34**	**84.04**	**82.46**
Flowers102	PTNL	**85.75**	76.54	73.11	56.22	36.82	**84.58**	77.44	60.94	36.17
	DP-SEP	84.47	**82.93**	**78.88**	**75.86**	**70.20**	81.90	**81.30**	**76.96**	**69.73**
Food101	PTNL	77.88	77.49	74.15	69.39	61.83	77.36	75.67	72.58	61.73
	DP-SEP	**78.46**	**78.43**	**78.01**	**77.75**	**77.28**	**78.34**	**78.12**	**77.97**	**77.52**
DTD	PTNL	**61.48**	**57.70**	50.61	43.46	35.80	57.78	53.27	44.15	30.15
	DP-SEP	59.27	56.54	**52.73**	**49.52**	**43.14**	**58.87**	**55.80**	**50.72**	**41.90**
UCF101	PTNL	**73.11**	70.29	64.63	59.75	52.31	71.58	66.66	60.78	49.58
	DP-SEP	72.97	**72.25**	**70.61**	**68.12**	**64.57**	**73.18**	**71.81**	**69.53**	**65.82**

Table 1: Performance Comparison under Synthetic Noise.

and instance-dependent (inst.) noise settings, achieving higher accuracy at most configurations. Notably, DP-SEP demonstrates significant improvements over PTNL in high-noise (in particular, achieves 33.11% accuracy gain under 80% sym. noise of OxfordPets and +16.24% under 80% inst. of UCF101), reaching an impressive average gain of 11.17% and 16.98% under 65% and 80% sym. noise, respectively. In lower noise (12.5% and 25%), performance gap between the two methods narrows, yet DP-SEP consistently outperforms PTNL across noise conditions, highlighting the robustness and stability of it. DP-SEP consistently maintains a significant advantage, particularly in challenging environments, demonstrating its effectiveness in mitigating the impact of incorrect annotations on complex visual features.

Performance under Real Noise: Performance on real noise benchmarks is shown in Table 2. Following standard practices in noisy label literature, we evaluate our models on the test sets of WebVision and ILSVRC12 [13] and report Top-1 and Top-5 accuracy. We use the test set of Food101 [3] for Food101N and report Top-1 accuracy. Consistent with synthetic noise, DP-SEP achieves superior Top-1 accuracy across all test settings. Notably, our method achieves 75.88% Top-1 accuracy on WebVision (2.03% improvement) and a 2.9% gain on Food101N. Given real-world noise distributions are often more complex and irregular than synthetic ones, these results provide a strong evidence of the effectiveness of DP-SEP.

4.2 Analysis

4.2.1 Robustness and Generalization.

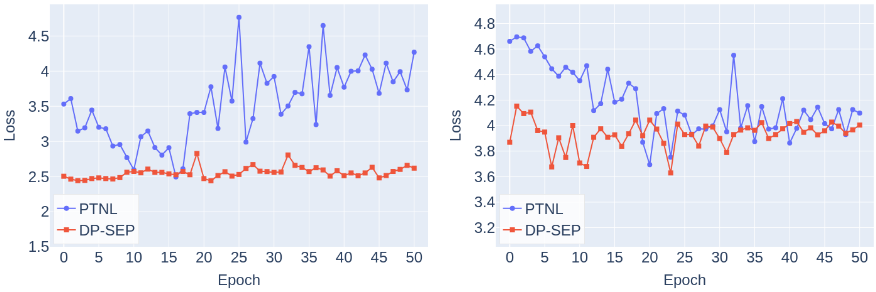
We evaluate **cross-dataset generalization** performance of DP-SEP and PTNL in Table 3 by training on ImageNet under various noise settings and validating on other datasets. DP-SEP consistently shows greater stability and adaptability, especially in high-noise settings. Notably, on UCF101, PTNL shows slight advantage in low-noise settings, however, the difference is minimal. This could be attributed to the unique nature of UCF101 as an action recognition dataset, which may respond differently compared to other classification datasets.

Method	WebVision Top-1	WebVision Top-5	ILSVRC12 Top-1	ILSVRC12 Top-5	Food101N
PTNL	73.55	95.95	73.29	**96.32**	61.61
DP-SEP	**75.58**	**96.17**	**75.23**	96.24	**64.51**

Table 2: Performance Comparison under Real Noise.

Dataset	Method	Symmetric 12.5%	25%	50%	65%	80%	Instance 20%	40%	60%	80%
OxfordPets	PTNL	79.78	79.39	79.10	77.92	74.73	79.37	**77.11**	78.22	71.93
	DP-SEP	**80.24**	**80.29**	**79.75**	**79.91**	**81.06**	**79.64**	76.48	**80.40**	**81.25**
Food101	PTNL	74.80	73.82	71.82	70.52	68.75	73.45	74.16	71.90	66.80
	DP-SEP	**75.45**	**75.07**	**75.18**	**75.71**	**76.71**	**74.20**	**74.44**	**76.13**	**76.59**
DTD	PTNL	34.84	31.01	34.15	32.07	29.79	34.57	35.43	34.79	30.27
	DP-SEP	**36.49**	**36.70**	**38.30**	**36.91**	**40.11**	**38.56**	**36.86**	**36.28**	**39.95**
UCF101	PTNL	**58.95**	58.52	**58.31**	54.37	50.91	**60.53**	**57.94**	55.17	52.26
	DP-SEP	57.63	**58.68**	57.78	**59.93**	**60.19**	58.47	57.57	**59.82**	**60.06**

Table 3: Cross-Dataset Generalization - Training on ImageNet and Testing on different datasets.



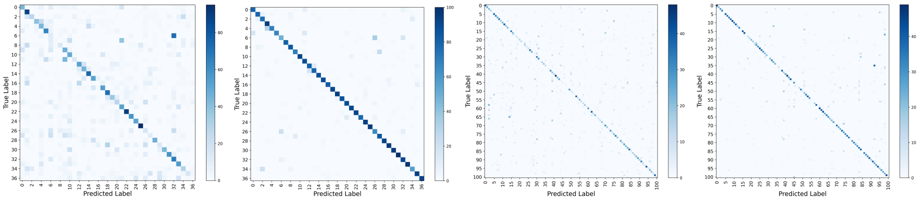
(a) OxfordPets (b) UCF101

Figure 2: Loss Dynamics of PTNL and DP-SEP using OxfordPets and UCF101 (Sym. 80% noise).

We further validate DP-SEP on different **context lengths, severe noise settings, varying few-shots and different backbones** (Detailed validation results in Supplement Section B). DP-SEP demonstrates robust adaptability and resilience across experiments and prove highly effective even under extreme noise and low-shot settings.

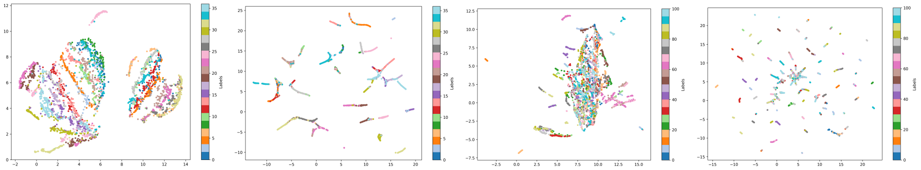
4.2.2 Model Behaviour

Loss Dynamics: We depict training loss of PTNL (blue) and DP-SEP (red) on Sym. 80% noise of OxfordPets and UCF101 in Figure 2. PTNL exhibits extreme fluctuations, especially in later stages of training, likely due to strong repetition of confidence in incorrect labels. In contrast, DP-SEP maintains a much steadier loss, indicating robustness in high-noise environments. DP prompts semantically strengthen the perception of difference between correct and incorrect descriptions, reducing oscillations caused by noisy labels. Additionally, SEP prevents overconfidence in unreliable labels, further promoting more robust convergence.



(a) OxfordPets-PTNL (b) OxfordPets-DP-SEP (c) UCF101-PTNL (d) UCF101-DE-SEP

Figure 3: Visualizing confusions of PTNL and DP-SEP on Inst. 60% noise.



(a) OxfordPets-PTNL (b) OxfordPets-DP-SEP (c) UCF101-PTNL (d) UCF101-DP-SEP

Figure 4: Feature visualization of PTNL and DP-SEP using UMAP (Inst. 60% noise).

Visualizing Confusions: We visualize confusion matrices of PTNL and DP-SEP on Inst. 60% noise of OxfordPets and UCF101 in Figure 3 (More visualizations in Supplement Section C). PTNL exhibits darker off-diagonal blocks, indicating higher misclassification rate and weaker class discrimination. In contrast, DP-SEP shows a more concentrated diagonal, signifying stronger accuracy and stability. Higher color intensity along the diagonal further confirms that DP-SEP consistently makes correct predictions across most categories, effectively mitigating the impact of label noise.

Feature Visualizations: We present UMAP [22] visualizations (only text encoders as image encoders are frozen) on Inst. 60% noise of OxfordPets and UCF101 in Figure 4 (and Supplement Section D). DP-SEP exhibits a well-separated and dispersed distribution, indicating a more distinct and structured feature space. In contrast, PTNL shows overlapping and entangled features, suggesting weaker class separability. These visualizations highlight the ability of DP-SEP to learn more discriminative representations, even under high-noise conditions.

4.2.3 Ablation Study

We analyze the contribution of each component of DP-SEP in Table 4. The most significant improvement comes from DP prompts (PAR Factor used directly with CE loss), particularly in high-noise, which leads to a 20% and 30% increase in accuracy under 65% and 80% noise, respectively. DP prompts effectively mitigate the confusion caused by incorrect labels. When SEP (with perturbation) is applied, the model shows +2-3% over PTNL under 50%, 65% and 80% noise, indicating its effectiveness in suppressing overconfidence in unreliable labels. When both components are combined, performance is further enhanced across all noise ratios, reflecting the synergistic effect of DP and SEP. DP prompts reduce the semantic impact of noisy labels, while SEP prevent overly trusting erroneous labels. This combination counteracts the bias caused by incorrect labels from multiple aspects. Ablations on PAR Objectives and Hyperparameter Sensitivity can be found in Supplement Section E.

Method	12.5%	25%	50%	65%	80%
PTNL (Baseline)	83.60	81.86	74.70	64.81	50.43
DP	85.72	84.33	83.97	84.06	82.91
SEP w/o perturbation	83.32	81.20	74.46	66.59	51.92
SEP	84.36	81.93	76.81	68.06	53.26
DP-SEP	**86.38**	**85.13**	**85.08**	**84.94**	**83.54**

Table 4: Ablation Study using OxfordPets on Sym. noise settings.

4.2.4 Limitations

DP-SEP exhibits certain limitations: (1) Despite significant improvements, DP-SEP could further benefit from integrating noise correction strategies to enhance noise mitigation. (2) Negative prompts though effective, currently operates at a global level. Investigating context-aware, adaptive strategies could unlock new capabilities.

5 Conclusions

As VLMs are increasingly becoming central to several applications, their sensitivity to label noise poses a significant barrier. In this regard, we focus on addressing the degradation of VLMs under high and complex noise conditions. We propose DP-SEP, a framework revolving around two core components: (1) Dual Polarity Prompt constructs the discriminative relationship between *correct description* and *wrong description* in the semantic space, limiting the model's over-reliance on labels (possibly wrong); (2) Stochastic Entropy Perturbation uses randomness to break the network's possibility of absolute confidence in incorrect samples. We conduct extensive experiments on multiple synthetic and real-world noise environments, and the results show that DP-SEP maintains robust discrimination capabilities under different noise configurations, far exceeding existing baselines. Based on these systematic experiments and in-depth analysis, we believe that DP-SEP not only expands the applicability of prompt tuning in noisy scenarios, but also brings new breakthroughs to VLMs in high-noise and complex environments. Developing prompting schedules based on input semantics and model behaviour and gaining deeper understanding of the optimization landscape are potential future directions.

Acknowledgments

This work was partially funded by the EU project MUSAE (No. 01070421), 2021-SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220-VET-000089434, Erasmus+ EU), DeepSense (ACE053/22/000029, ACCIÓ), DeepFoodVol (AEI-MICINN, PDC2022-133642-I00) and IDEATE (AEI-MICINN, PID-2022-141566NB-I00). C. Hu acknowledges China Scholarship Council (CSC) Fellowship (Ministry of Education, China, No. 202208410100). B. Nagarajan acknowledges AI4S fellowship within the "Generación D" initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by NextGenerationEU through PRTR.

References

[1] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36:80396–80413, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014.

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[5] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023.

[6] Li Fei-Fei, Rob Fergus, and Pietro 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*, pages 178–178. IEEE, 2004.

[7] Yuncheng Guo and Xiaodong Gu. Joapr: Cleaning the lens of prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28695–28705, 2024.

[8] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[9] Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15670–15680, 2023.

[10] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

[11] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023.

[12] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[14] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[15] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5456, 2018.

[16] Ming Li, Jike Zhong, Chenxin Li, Liuzhuozheng Li, Nie Lin, and Masashi Sugiyama. Vision-language model fine-tuning via simple parameter-efficient modification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14394–14410, 2024.

[17] W Li, L Wang, W Li, E Agustsson, and L Van Gool. Webvision database: Visual learning and understanding from web data. arxiv 2017. *arXiv preprint arXiv:1708.02862*.

[18] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[19] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[20] Chengcheng Ma, Yang Liu, Jiankang Deng, Lingxi Xie, Weiming Dong, and Changsheng Xu. Understanding and mitigating overfitting in prompt tuning for vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4616–4629, 2023.

[21] Ségolène Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, and Ismail Ben Ayed. Transductive zero-shot and few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28816–28826, 2024.

[22] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[23] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[25] Jinyoung Park, Juyeon Ko, and Hyunwoo J Kim. Prompt learning via meta-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26940–26950, 2024.

[26] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[27] Qi Qian and Juhua Hu. Online zero-shot classification with clip. In *European Conference on Computer Vision*, pages 462–477. Springer, 2024.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[29] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[30] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[31] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024.

[32] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17542–17552, 2024.

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[34] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[35] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024.

[36] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13019–13029, 2024.

[37] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[38] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[39] Qitong Wang, Tang Li, Kien X Nguyen, and Xi Peng. Beyond accuracy: On the effects of fine-tuning towards vision-language model's prediction rationality. *arXiv preprint arXiv:2412.13333*, 2024.

[40] Tong Wei, Hao-Tian Li, ChunShu Li, Jiang-Xin Shi, Yu-Feng Li, and Min-Ling Zhang. Vision-language models are strong noisy label detectors. *Advances in Neural Information Processing Systems*, 37:58154–58173, 2024.

[41] Cheng-En Wu, Yu Tian, Haichao Yu, Heng Wang, Pedro Morgado, Yu Hen Hu, and Linjie Yang. Why is prompt tuning for vision-language models robust to noisy labels? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15488–15497, 2023.

[42] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33:7597–7610, 2020.

[43] Jialu Xing, Jianping Liu, Jian Wang, Lulu Sun, Xi Chen, Xunxun Gu, and Yingfei Wang. A survey of efficient fine-tuning methods for vision-language models—prompt and adapter. *Computers & Graphics*, 119:103885, 2024.

[44] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[45] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

[46] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[47] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12924–12933, 2024.

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

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[50] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11175–11185, 2023.