NLPrompt: Noise-Label Prompt Learning for Vision-Language Models

Bikang Pan[1,†] Qun Li[1,†] Xiaoying Tang[2] Wei Huang[3] Zhen Fang[4] Feng Liu[5]
Jingya Wang[1] Jingyi Yu[1] Ye Shi[1,*]
[1]ShanghaiTech University, Shanghai, China
[2]The Chinese University of Hong Kong, Shenzhen, China
[3]RIKEN Center for Advanced Intelligence Project, Japan
[4]University of Technology Sydney, Australia [5]University of Melbourne, Australia
{panbk2023,liqun2024,wangjingya,yujingyi,shiye}@shanghaitech.edu.cn,
tangxiaoying@cuhk.edu.cn,wei.huang.vr@riken.jp,
zhen.fang@uts.edu.au,feng.liu1@unimelb.edu.au
https://github.com/qunovo/NLPrompt

Abstract

The emergence of vision-language foundation models, such as CLIP, has revolutionized image-text representation, enabling a broad range of applications via prompt learning. Despite its promise, real-world datasets often contain noisy labels that can degrade prompt learning performance. In this paper, we demonstrate that using mean absolute error (MAE) loss in prompt learning, named PromptMAE, significantly enhances robustness against noisy labels while maintaining high accuracy. Though MAE is straightforward and recognized for its robustness, it is rarely used in noisy-label learning due to its slow convergence and poor performance outside prompt learning scenarios. To elucidate the robustness of PromptMAE, we leverage feature learning theory to show that MAE can suppress the influence of noisy samples, thereby improving the signal-to-noise ratio and enhancing overall robustness. Additionally, we introduce PromptOT, a prompt-based optimal transport data purification method to enhance the robustness further. PromptOT employs text features in vision-language models as prototypes to construct an optimal transportation matrix. This matrix effectively partitions datasets into clean and noisy subsets, allowing for the application of cross-entropy loss to the clean subset and MAE loss to the noisy subset. Our Noise-Label Prompt Learning method, named NLPrompt, offers a simple and efficient approach that leverages the expressive representations and precise alignment capabilities of vision-language models for robust prompt learning. We validate NLPrompt through extensive experiments across various noise settings, demonstrating significant performance improvements.

1. Introduction

The advent of vision-language foundation models, such as CLIP [46], has revolutionized how images and their textual descriptions are represented, providing a unified perspective for both modalities. In these models, images are typically aligned with sentences like “A photo of a $\langle$CLS$\rangle$”, thereby facilitating the efficient handling of various tasks. Given the sensitivity of handcrafted text in descriptions, prompt learning has emerged as a crucial method for fine-tuning these vision-language models. Prompt learning involves updating a learnable text prompt through back-propagation [8, 11, 35, 60, 61], offering a lightweight solution due to the relatively small number of parameters involved, often just several thousand. This adaptability ensures rapid tuning for specific tasks.

Nevertheless, real-world applications often face the challenge of dealing with noisy labels in annotated datasets, necessitating robust learning solutions. Prior work [53] illustrates that prompt tuning is more resilient to noisy labels compared to other fine-tuning paradigms such as adapter tuning. Despite this, prompt tuning remains vulnerable to overfitting when trained with cross-entropy loss under noisy conditions. Therefore, enhancing the robustness of prompt tuning in noisy environments remains a crucial issue.

In the realm of noisy label learning, mean absolute error (MAE) has been identified as a robust loss function within the traditional training paradigm [17]. However, MAE often suffers from slow convergence and poor performance during training, making it seldom employed as a classification loss in noise-label learning. Nevertheless, our investigation reveals an interesting phenomenon: *employing MAE loss in Prompt learning (PromptMAE) notably enhances robustness while maintaining high accuracy compared to tra-*

[*]Corresponding author. † Equal contribution.

ditional cross-entropy loss. As demonstrated in Figure 1, MAE exhibits strong accuracy and fast convergence even in the presence of substantial noise.

To elucidate the robustness of PromptMAE, we leverage feature learning theory [1, 3, 41], which categorizes latent representations into task-relevant and task-irrelevant components. By analyzing the optimization dynamics of these features with gradient-descent-based training, we can gain valuable insights into convergence and generalization. To this end, we find that robust prompt learning is achieved when task-relevant features dominate. Our analysis indicates that PromptMAE can suppress the influence of noisy samples, thereby enhancing robustness in prompt learning for vision-language models.

A standard approach in noisy label learning is the employment of sample selection techniques [9, 14, 20, 36, 44] to clean the dataset and thus improve performance under noisy conditions. For example, optimal transport (OT)-based sample selection methods [14, 54] utilize randomly initialized prototypes to compute the optimal transportation matrix from image features to these prototypes, considering the similarity between features and prototypes as a cost matrix. As these methods were not originally designed for prompt learning, their direct applicability may be limited. We aim to harness the inherent alignment in vision-language foundation models to refine the data purification process.

In this paper, we introduce PromptOT, a prompt-based optimal transport data purification method, designed to enhance the robustness of prompt learning in vision-language foundation models. PromptOT leverages the text features as prototypes for the transportation matrix, facilitating robust prompt learning by partitioning the dataset into clean and noisy subsets. Recognizing that cross-entropy (CE) loss generally outperforms MAE on clean datasets, we apply MAE loss to train the noisy subset and cross-entropy loss to train the clean subset. This dual strategy, supported by PromptOT purification, harmonizes the strengths of both MAE and CE loss under varying noisy conditions. Our comprehensive method, named NLPrompt, leverages the expressive representation and alignment capabilities of vision-language models, offering a simple and efficient solution for robust prompt learning in the presence of noisy labels. In summary, our contributions are threefold:

- We discover that a simple MAE loss significantly improves the robustness of prompt learning on noisy datasets. Utilizing feature learning theory, we theoretically demonstrate how PromptMAE reduces the impact of noisy samples, enhancing overall robustness.
- We introduce NLPrompt, a robust prompt learning method that uses a simple MAE loss with PromptOT-based data purification to handle noisy labels. NLPrompt efficiently exploits the expressive representation and precise alignment capabilities of vision-language foundation models for robust prompt learning.
- We validate the effectiveness of NLPrompt through extensive experiments across datasets with varied noise conditions, consistently showing significant performance improvements.

2. Related Work

2.1. Prompt Learning in Vision-Language Models

Prompt learning, which began in natural language processing, has now extended into the realm of vision-language models. A notable example is the CLIP model [46], which initially relied on hand-crafted prompts. Recent advancements, however, have shifted focus towards learning prompts in a continuous embedding space. Innovations such as CoOp [60] have enhanced the CLIP model by integrating continuous prompt vectors, fostering a wave of research dedicated to optimizing prompt learning and paving the way for further exploration. In addition to CoOp, CoCoOp [61] utilizes a neural network to generate input-specific context tokens that adapt the prompts based on each image, thereby improving generalization to unseen classes. ProGrad [62] regularizes the soft prompt updates by aligning their gradients with the general knowledge provided by the original prompt. MaPLe [29] introduces branch-aware hierarchical prompts that address both language and vision branches. TPT (Test-Time Prompt Tuning) [47] explores prompt tuning without additional training samples by augmenting the input image into various views and training the learnable prompts to generate consistent responses across these different views.

2.2. Learning with Noisy Labels

Mislabeled data can lead to deep neural networks overfitting to noisy labels. To address the issue of learning from noisy labels, previous researchers have proposed various methods, including robust network architectures [33, 57], robust regularization techniques [22, 38, 56], robust loss functions [15, 17, 37], correction of loss via estimation matrices [5, 55, 58], and sample selection and meta-learning approaches [36, 44, 49].

The study of prompt learning with noise labels is currently in its nascent stage. A pioneering work by Wu et al. [53] demonstrated that prompt learning is more robust than other parameter-efficient fine-tuning methods, such as adapters. Subsequently, JoAPR [19] uses a Gaussian mixture model with joint adaptive thresholds to differentiate between clean and noisy data. It corrects labels by combining the results of data augmentations with mixup loss, and then

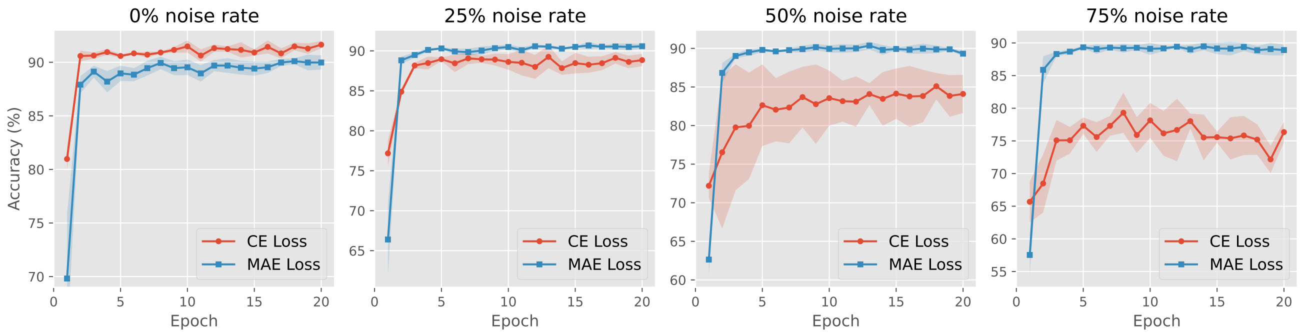

Figure 1. The performance of training with MAE loss and CE loss in prompt learning on Caltech101 dataset.

retrains the model using this refined data. However, this approach does not fully leverage the benefits of prompt learning. In contrast, our study shows that using a simple MAE loss in prompt learning can boost the capability for handling noisy data. Additionally, we incorporate prompt-based optimal transport to further purify the noise samples.

2.3. Feature Learning Theory

To further understand how noisy label learning affects prompt learning, we leverage feature learning theory to analyze the learning process. Feature learning theory [1, 3, 23–25, 28, 51, 63] categorizes latent features into task-relevant and task-irrelevant components, expressing the trainable weights as a combination of these feature types. From this perspective, feature learning analyzes the coefficients of these features to gain insight into learning dynamics. Beyond its application in traditional learning paradigms, prompt learning can also be explained by feature learning theory [41]. In this paper, we adopt feature learning theory to demonstrate the robustness of MAE in prompt learning.

3. Preliminary

Notation. In our work, vectors are represented by lowercase bold letters, matrices by uppercase bold letters, and scalars by regular, non-bold letters. The ℓ_2-norm of a vector $\mathbf{v}$ is denoted as $|\mathbf{v}|_2$. For matrices, the spectral norm of $\mathbf{A}$ is indicated by $|\mathbf{A}|_2$, and the Frobenius norm by $|\mathbf{A}|_F$. The indicator function is represented by $\mathbb{1}(\cdot)$. Finally, sequences of integers are represented as $[n] = \{1, 2, \ldots, n\}$, and sequences of elements, such as vectors, are similarly denoted as $\mathbf{v}_{[n]} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$.

Prompt Learning. In this section, we demonstrate how to fine-tune a learnable text prompt within a vision-language pre-trained model. We focus on a classification task, where we have an image $\mathbf{x}$ that we aim to classify into the correct ground truth class $y \in [C]$, with C representing the total number of classes. From the vision-language pre-trained model, we expect the latent spaces of the text encoder and image encoder to be aligned. This alignment ensures that, when different prompts are input, the text feature generated by the correct prompt will have the highest similarity with the image feature. In this setup, we input a learnable prompt $\mathbf{p} \in \mathbb{R}^d$ along with a fixed class prompt $\mathbf{p}_c \in \{\mathbf{p}_1, \ldots, \mathbf{p}_C\}$, where each $\mathbf{p}_c \in \mathbb{R}^d$ represents a specific class, into the text encoder h. Here, d denotes the dimensionality of the prompts. By incorporating the learnable prompt $\mathbf{p}$, we generate the text feature for class c as $\mathbf{h}_c = h(\mathbf{p}, \mathbf{p}_c) \in \mathbb{R}^m$. Meanwhile, the image feature $\mathbf{g}$ is produced by the image encoder g as $\mathbf{g} = g(\mathbf{x}) \in \mathbb{R}^m$. We define the similarity function between the image feature $\mathbf{g}$ and the text feature $\mathbf{h}_c$ as $\boldsymbol{\rho} = \mathrm{sim}(\mathbf{g}, \mathbf{h}_c) \in \mathbb{R}^c$. The training process follows the structure of traditional classification tasks, with an objective loss $\ell(\boldsymbol{\rho}, \mathbf{e}_y)$ that measures the distance between the similarity vector $\boldsymbol{\rho}$ and the true label y. Here, ℓ represents the loss function quantifying the distance between two vectors, and $\mathbf{e}_y$ is the one-hot vector associated with the ground truth label y.

Optimal Transport. Optimal transport (OT) is a constrained optimization problem that seeks to determine the optimal coupling matrix that maps one probability distribution to another while minimizing the total cost. Given the marginal distributions $\boldsymbol{\alpha} \in \mathbb{R}^n$, $\boldsymbol{\beta} \in \mathbb{R}^m$, and the cost matrix $\mathbf{C} \in \mathbb{R}^{n \times m}$, the classical OT problem is formulated as follows:

$$\begin{aligned} \min_{\mathbf{Q} \in \mathbb{R}_+^{n \times m}} \quad & \langle \mathbf{C}, \mathbf{Q} \rangle \\ \text{s.t.} \quad & \mathbf{Q}\mathbb{1}_m = \boldsymbol{\alpha}, \ \mathbf{Q}^\top \mathbb{1}_n = \boldsymbol{\beta}. \end{aligned} \tag{1}$$

This problem is a linear programming task, which becomes computationally expensive as the problem scale increases. To address this, Sinkhorn [12] proposed adding an entropic regularization, which allows for a closed-form solution and provides a "lightspeed" algorithm that only requires iterative scaling of the transportation matrix. The entropic regularized formulation is given as follows:

$$\begin{aligned} \min_{\mathbf{Q}\in\mathbb{R}_+^{n\times m}} \quad & \langle \mathbf{C}, \mathbf{Q}\rangle - \epsilon H(\mathbf{Q}) \\ \text{s.t.} \quad & \mathbf{Q}\mathbb{1}_m = \boldsymbol{\alpha},\ \mathbf{Q}^\top \mathbb{1}_n = \boldsymbol{\beta}, \end{aligned} \tag{2}$$

where $H(\mathbf{Q}) = \sum_{i,j} Q_{ij} \left(\log Q_{ij} - 1\right)$ and $\epsilon \geq 0$ is the coefficient that controls the regularization term. In previous work, OT has been formulated as a pseudo-labeling technique for a range of machine learning tasks, including class-imbalanced learning [18, 50], semi-supervised learning [32, 39], clustering [2, 4, 16], domain adaptation [6, 59], label refinery [7, 14, 50, 54], and others. Unlike prediction-based pseudo-labeling [48], OT-based pseudo-labeling optimizes the mapping samples to class centroids, while considering the global structure of the sample distribution in terms of marginal constraints instead of per-sample predictions.

4. Theoretical Analysis for the Robustness of PromptMAE

In prompt learning with noisy labels, the ground truth class y is flipped to a different noisy label class with a certain probability. As shown in Figure 1, we compared the performance of the original CoOp using cross-entropy (CE) loss and mean absolute error (MAE) loss. We observed that as the noise level increased in the datasets, the performance using CE loss significantly dropped, while the mean absolute error loss showed negligible change.

Basic Settings. To explain this phenomenon, we apply feature learning theory to characterize the mechanism behind the robustness of MAE in the context of prompt learning. In our analysis, the objective is to classify the image $\mathbf{x}$ into its true class label y. In this theoretical analysis, we focus on a binary classification scenario where the class label $y_i \in \{+1, -1\}$. We also assume that the latent spaces of both the text encoder and the image encoder in the vision-language pre-trained model are well aligned. The latent feature space is assumed to consist of both task-relevant features, denoted as $\boldsymbol{\mu} \in \mathbb{R}^m$, and task-irrelevant features $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_L \in \mathbb{R}^m$, where L represents the number of task-irrelevant features and the dimension of the latent space is m. For simplicity, we assume these features are orthogonal to each other.

Text Encoder. Adopting a setup similar to [52], a learnable prompt $\mathbf{p}$ and a fixed class prompt $\mathbf{p}_c$ are fed into the text encoder h:

$$\begin{aligned} \mathbf{h}_c &= h(\mathbf{p}, \mathbf{p}_c) \\ &= \sigma(\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_c) - \sigma(-\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_c), \end{aligned} \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{m\times d}$ is the weight matrix, and $\mathbf{p}_c \in \mathbb{R}^d$ is the prompt associated with class c. To examine the properties of the text encoder in the pre-trained model, we follow [26, 41] and set the weight matrix $\mathbf{W}$ to be:

$$\mathbf{W} = \left[\boldsymbol{\mu}, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_L\right]^\top. \tag{4}$$

Image Encoder. Let us consider the image network, represented as $\mathbf{g}_i = g(\mathbf{x}_i) \in \mathbb{R}^m$. Due to that we assume the image encoder g aligns the feature space of the text encoder h. As a result, the image feature generated by data $\mathbf{x}_i$ in client i can be expressed as:

$$\mathbf{g}_i = g(\mathbf{x}_i) = \left[y_i, x_{i,1}, \cdots, x_{i,L}\right]^\top, \tag{5}$$

where $x_{i,l} \sim \mathcal{N}(0, \sigma_p^2), \forall l \in [L]$ represents the coefficient of task-irrelevant terms in the data, and σ_p^2 is the variance. The similarity score between an image $\mathbf{x}_i$ and class y_i is given by $\text{sim}(\mathbf{g}_i, \mathbf{h}_i) = \langle \mathbf{g}_i, \mathbf{h}_i\rangle$. To compute the probability, we first pass the logits of the similarity vector through the softmax function:

$$s_i(\mathbf{p}) = \text{SOFTMAX}(\text{sim}(\mathbf{g}_i, \mathbf{h})). \tag{6}$$

The CE loss and MAE loss are defined as:

$$\ell_{\text{CE}}(\mathbf{s}_i, \mathbf{y}_i) = \sum_{c=1}^{C} -y_{i,c} \log s_{i,c}, \tag{7}$$

$$\ell_{\text{MAE}}(\mathbf{s}_i, \mathbf{y}_i) = \sum_{c=1}^{C} |y_{i,c} - s_{i,c}|. \tag{8}$$

Noisy Label Modeling. Here, we introduce a model for label noise. We assume that the label noise follows a Rademacher random variable. Specifically, the noisy label $\tilde{y}$ is generated from the ground truth label y with probability $p \leq 1/2$. That is, $\mathbb{P}[\tilde{y} = -y] = p$ and $\mathbb{P}[\tilde{y} = y] = 1 - p$. For the purpose of analysis, we divide the entire dataset into two subsets: the clean dataset $S_+ = \{i \mid \tilde{y}_i = y_i\}$ and the noisy dataset $S_- = \{i \mid \tilde{y}_i = -y_i\}$.

Feature Representation. Under feature learning theory, the weight of the prompt can be decomposed into a combination of task-relevant features and task-irrelevant features. We present the following feature representation lemma [41]:

Lemma 4.1. *At the t-th iteration, the learnable prompt $\mathbf{p}^{(t)}$ can be rewritten as a linear combination of the features and the prompt initialization:*

$$\mathbf{p}^{(t)} = \alpha^{(t)}\mathbf{p}^{(0)} + \beta^{(t)}||\boldsymbol{\mu}||_2^{-2}\boldsymbol{\mu} + \sum_{l=1}^{L} \phi_l^{(t)}||\boldsymbol{\xi}_l||_2^{-2}\boldsymbol{\xi}_l,$$

where $\alpha^{(t)}$ are the coefficients of the initialization, $\beta^{(t)}$ and $\phi_l^{(t)}$ are the coefficient of the task-relevant features and task-irrelevant features, respectively.

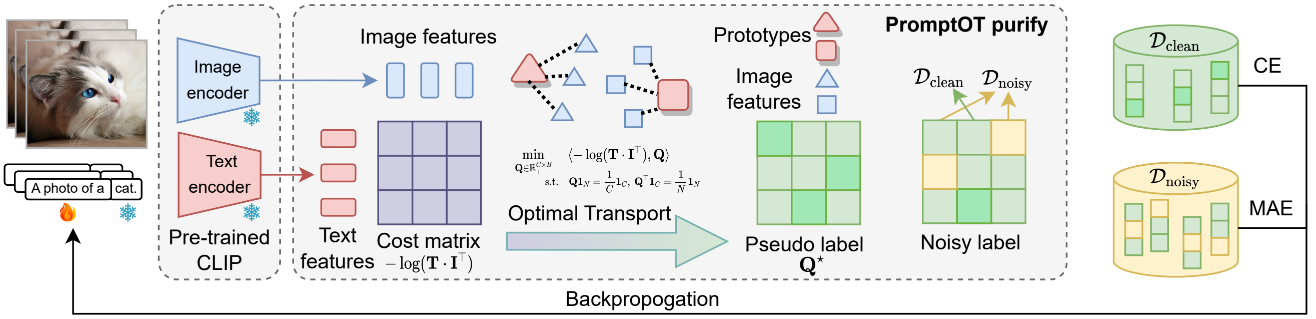
Image encoder
Text encoder
A photo of a
cat.
Pre-trained CLIP
Image features
Text features
Cost matrix
$-\log(\mathbf{T}\cdot\mathbf{I}^{\top})$
$\min_{\mathbf{Q}\in\mathbb{R}_{+}^{C\times B}} \langle -\log(\mathbf{T}\cdot\mathbf{I}^{\top}), \mathbf{Q}\rangle$
s.t. $\mathbf{Q}\mathbb{1}_N = \frac{1}{C}\mathbb{1}_C,\ \mathbf{Q}^{\top}\mathbb{1}_C = \frac{1}{N}\mathbb{1}_N$
Optimal Transport
Prototypes
Image features
Pseudo label
$\mathbf{Q}^{\star}$
PromptOT purify
$\mathcal{D}_{\text{clean}}$
$\mathcal{D}_{\text{noisy}}$
Noisy label
$\mathcal{D}_{\text{clean}}$
CE
$\mathcal{D}_{\text{noisy}}$
MAE
Backpropogation

Figure 2. The framework of our NLPrompt. We utilize the text representation to initialize prompt-based OT, which separates the dataset into clean and noisy subsets. NLPrompt harmonizes the advantage of MAE loss and CE loss. The former is more robust on the noisy dataset while the latter performs better on the clean dataset.

Since the learnable prompt can be expressed as a linear combination of the features, we can analyze the dynamics of these coefficients to understand the learning progress of the prompts. The normalization factor, such as $||\boldsymbol{\mu}||_2^{-2}$, ensures that the coefficients are comparable to the inner product of the prompt and the features, i.e., $\beta^{(t)} \approx \langle \mathbf{p}^{(t)}, \boldsymbol{\mu}\rangle$.

Robustness of PromptMAE. Building on the previous setup, we now examine how label flipping noise affects the learning dynamics of the coefficients. We will show that, the PromptMAE loss can enhance task-relevant coefficients for clean samples and help mitigate the degradation of task-relevant coefficients for noisy samples.

By analyzing the dynamics of these coefficients, we gain deeper insights into the learning process. As shown in [41], the performance of prompt fine-tuning can be evaluated based on the ratio between task-relevant and task-irrelevant coefficients. To illustrate that MAE produces more robust results, we present the following theorem:

Theorem 4.2. *With high probability at test $1-d^{-1}$, the test loss $\ell_{\mathcal{D}}$ for the prompt trained by MAE is lower than the prompt trained by CE, i.e., $\ell_{\mathcal{D}}(\mathbf{p}_{MAE}) \leq \ell_{\mathcal{D}}(\mathbf{p}_{CE})$. Here, $\mathbf{p}_{MAE}$ and $\mathbf{p}_{CE}$ refer to the text prompt trained using MAE loss and CE loss, respectively.*

The proof is provided in the Appendix. From this result, we observe that under a noisy dataset, the MAE loss demonstrates greater robustness.

5. Methodology

In this section, we introduce our NLPrompt algorithm and explain how we utilize the OT problem for data purification. NLPrompt harmonizes the advantage of MAE loss and CE loss. Theorem 4.2 has shown that MAE loss is more robust on the noisy dataset.

PromptOT Purification. Here, we utilize OT to generate pseudo-labels for data purification. Traditionally, the OT-based pseudo-labeling method starts with the random initialization of prototypes, and pseudo-labels are then derived from the similarity between images and these prototypes. However, in the context of prompt learning with vision-language foundation models, where the latent space is aligned, the randomly initialized prototypes can be replaced with text features generated by prompts via text encoders. The semantic information embedded in these text features provides a strong foundation for initialization.

Specifically, as outlined in Equation (2), the OT problem involves solving for a transportation matrix based on a given cost matrix, while preserving the marginal distributions. The similarity between the prototypes and the image features is calculated, and the negative logarithm of the resulting similarity matrix is used as the cost matrix. Due to the marginal distribution constraints, the columns of the OT matrix are normalized, and this matrix is then used as pseudo-labels for the images.

The calculation process in NLPrompt is outlined below. For images in dataset $\{\mathbf{x}_i\}_{i=1}^{N}$, we first use the pre-trained image encoder of CLIP to generate an image feature matrix $\mathbf{I} \in \mathbb{R}^{N\times d}$, where d represents the dimension of the latent space. Additionally, given the set of classes, we generate prompts corresponding to these classes and pass them to the pre-trained text encoder of CLIP to create a text feature matrix $\mathbf{T} \in \mathbb{R}^{C\times d}$, where C is the number of classes. Next, we calculate the similarity matrix $\mathbf{T}\cdot\mathbf{I}^{\top}$. The negative logarithm of this similarity matrix is then used as the cost matrix in the OT problem, with uniform marginal distributions across both the samples and the classes. The OT problem to be solved is given as follows:

$$\min_{\mathbf{Q}\in\mathbb{R}_{+}^{C\times N}} \quad \langle -\log(\mathbf{T}\cdot\mathbf{I}^{\top}), \mathbf{Q}\rangle \tag{9}$$
$$\text{s.t.} \quad \mathbf{Q}\mathbb{1}_N = \frac{1}{C}\mathbb{1}_C,\ \mathbf{Q}^{\top}\mathbb{1}_C = \frac{1}{N}\mathbb{1}_N.$$

From this formulation, we obtain the OT matrix $\mathbf{Q}^{\star}$. We then apply the Argmax operation to each column of $\mathbf{Q}^{\star}$ to find the maximum value, i.e.,

$$\tilde{y}_i = \arg\max_j \mathbf{Q}_{ij}.$$

Harmonizing MAE and CE within NLPrompt. The pseudo-labels generated by PromptOT are used to purify the dataset into two subsets: the clean dataset $\mathcal{D}_{\text{clean}}$ and the noisy dataset $\mathcal{D}_{\text{noisy}}$, defined as follows:

$$\mathcal{D}_{\text{clean}} = \{i \mid \hat{y}_i = \tilde{y}_i\}, \quad \mathcal{D}_{\text{noisy}} = \{j \mid \hat{y}_j \neq \tilde{y}_j\}. \quad (10)$$

After the split, the two subsets are trained using different loss functions. For the clean dataset, we leverage the high performance of CE loss, while for the noisy dataset, we use MAE loss to enhance robustness. The harmonizing loss can be expressed as

$$\ell_{\text{NLPrompt}} = \sum_{i \in \mathcal{D}_{\text{clean}}} -\mathbf{y}_i^\top \log \mathbf{s}_i + \sum_{j \in \mathcal{D}_{\text{noisy}}} ||\mathbf{y}_j - \mathbf{s}_j||_1. \quad (11)$$

where $\mathbf{y}_i$ is the target label and $\mathbf{s}_i$ is the output similarity for i-th sample.

Remark. Our NLPrompt utilizes OT to harmonize CE and MAE. Unlike using generalized CE loss in noise-label prompt learning [53], our method fully exploits the advantages of prompt learning under vision-language foundation models. First, we utilize the text representation from prompt learning as a strong initial prototype. This allows our method to maintain global label consistency, setting it apart from other prediction-based methods. Additionally, we refine the dataset to take advantage of the robustness of mean absolute error, specifically for noisy samples, rather than treating both clean and noisy samples with the same loss. This flexibility not only enhances our model's robustness but also allows us to leverage the advantages of CE, leading to improved overall performance.

6. Experiments

In this section, we conduct comprehensive experiments to evaluate the performance of our method in noisy label scenarios, demonstrating the effectiveness of our method.

6.1. Datasets and Baselines

Datasets. To evaluate the performance of our method, we conduct experiments on seven synthetic noisy datasets: Caltech101 [13], DTD [10], EuroSAT [21], Flowers102 [40], OxfordPets [42], StanfordCars [31], and UCF101 [45]. These representative visual classification datasets are used to simulate datasets with limited samples. They cover a variety of tasks, including general object classification, texture classification, fine-grained classification, action recognition, and satellite imagery recognition. Since these datasets do not contain noisy labels by default, we manually generate noisy labels for them. In addition, we conduct experiments on a real-world noisy label dataset, Food101N [34], which inherently contains noisy labels and does not require manually synthesized noisy labels.

Baselines. We compare our NLPrompt method with three baselines : CoOp [60], CoOp+GCE [53], JoAPR [19]. The latter two methods are specifically designed to tackle label noise in prompt learning for prompt learning in vision-language pretrained Models.

6.2. Noise Setting

For these synthetic noisy datasets, we introduce two types of noise: symmetric noise (denoted as Sym) and asymmetric noise (denoted as Asym). We only flip the labels of the training set in these datasets while keeping the test set unchanged. For symmetric noise, the clean labels in the training set are randomly flipped to other labels with equal probability. This means that labels within the same class can be incorrectly mapped to multiple different classes. For asymmetric noise, the clean labels in the training set are only flipped to a unique neighboring label, with labels within the same class being mapped exclusively to their successor class. Due to its stronger structural nature, asymmetric noise has a more significant negative impact on model performance, making it a stricter robustness test to evaluate the model's stability and adaptability when confronted with label noise. The goal of learning with noisy labels is to train a robust model on the noisy training set and achieve high accuracy on the clean test set.

6.3. Implementation Details

In our experiments, we adopt the same setup as CoOp [60] and JoAPR [19] to ensure a fair comparison. We use the SGD optimizer with an initial learning rate of 0.002 and employ cosine annealing. Our model backbone is consistent with CoOp [60], based on the pre-trained CLIP model [46], utilizing either ResNet-50 or ViT-B/16 as the image encoder, with ResNet-50 as the default if not explicitly specified. We use a 63M parameter text transformer as the text encoder. The default number of training epochs is set to 200. Additionally, we employ 16 shared context tokens across all categories, with the class token placed at the end. We sample a 16-shot training set from each dataset and evaluate the model on the original test set. The reported experimental results are the averages of test accuracy from three runs with different seeds, with the highest accuracy highlighted in bold. All experiments were conducted using PyTorch [43] on a cluster equipped with NVIDIA A40 GPU. It is noted that our noise setting differs from that of JoAPR. Our setting is more challenging and practical. The differ-

Table 1. Performance metrics across various datasets and noise levels. (%)

Dataset	**Method**	**Noise Rate: Sym**						**Noise Rate: Asym**					
		12.5%	**25.0%**	**37.5%**	**50.0%**	**62.5%**	**75.0%**	**12.5%**	**25.0%**	**37.5%**	**50.0%**	**62.5%**	**75.0%**
Flowers102	CoOp	88.93	83.50	77.93	70.10	55.60	37.17	86.97	74.70	60.43	42.60	26.53	12.60
	GCE	88.80	88.33	86.73	84.07	78.37	70.37	88.40	86.37	80.33	69.93	61.50	39.23
	JoAPR	85.57	81.23	74.60	70.23	67.90	66.93	85.17	79.63	73.97	73.83	53.37	13.27
	NLPrompt	**93.87**	**92.57**	**92.73**	**89.90**	**84.77**	**76.80**	**93.80**	**93.40**	**91.77**	**81.10**	**73.63**	**55.33**
DTD	CoOp	56.00	49.57	43.30	34.37	27.83	17.27	55.60	47.75	38.07	29.63	20.53	11.70
	GCE	61.00	59.83	56.80	50.73	43.60	33.67	60.70	57.57	52.70	43.97	33.40	18.23
	JoAPR	58.07	57.70	56.33	53.03	48.05	29.90	52.40	56.63	53.10	48.93	40.20	28.26
	NLPrompt	**62.97**	**61.23**	**59.17**	**55.17**	**49.03**	**39.80**	**62.30**	**60.60**	**56.47**	**50.80**	**40.27**	**28.37**
EuroSAT	CoOp	76.50	69.23	61.67	52.33	37.63	26.70	76.00	66.27	53.83	41.17	28.00	17.43
	GCE	82.13	78.60	74.67	63.13	49.67	31.40	78.23	72.70	63.63	45.30	22.90	12.10
	JoAPR	75.13	61.10	60.90	63.63	38.97	27.33	69.37	67.30	59.40	47.60	33.93	17.50
	NLPrompt	**82.53**	**79.53**	**78.13**	**66.70**	**63.53**	**43.80**	**80.13**	**77.13**	**71.43**	**54.30**	**66.33**	**32.73**
OxfordPets	CoOp	76.50	66.73	60.33	47.03	35.77	24.60	76.10	66.20	52.53	38.73	26.63	14.90
	GCE	85.63	84.60	83.67	79.23	71.40	53.17	85.50	83.03	76.73	68.07	50.70	31.97
	JoAPR	84.00	83.26	83.20	83.10	82.40	**74.40**	82.90	83.40	79.07	75.84	52.74	43.57
	NLPrompt	**86.17**	**86.00**	**85.33**	**84.87**	**83.63**	70.77	**86.00**	**84.97**	**82.40**	**77.53**	**66.33**	**48.60**
StanfordCars	CoOp	66.20	59.70	53.40	45.90	35.67	22.90	65.77	57.13	46.23	33.73	22.37	12.80
	GCE	**69.70**	66.40	66.47	63.77	59.25	50.87	**70.00**	66.45	61.23	53.67	39.65	26.60
	JoAPR	68.60	66.30	62.83	56.67	48.50	39.40	66.47	61.70	51.50	42.03	30.80	22.97
	NLPrompt	69.37	**68.80**	**67.20**	**65.63**	**62.83**	**58.30**	69.77	**67.53**	**64.23**	**59.03**	**50.90**	**39.50**
UCF101	CoOp	69.03	63.40	58.23	49.73	40.83	26.30	67.23	58.07	46.47	34.43	23.67	13.17
	GCE	74.00	**73.63**	72.57	69.37	66.00	57.07	73.90	71.87	67.97	62.23	52.50	36.37
	JoAPR	72.83	71.17	70.37	67.63	65.30	57.67	72.07	69.80	64.10	59.17	56.07	47.46
	NLPrompt	**74.83**	73.40	**72.83**	**70.33**	**68.10**	**60.53**	**74.90**	**73.53**	**71.03**	**65.97**	**58.97**	**49.27**
Caltech101	CoOp	86.43	81.03	76.73	70.90	61.33	46.90	84.93	75.23	62.87	49.43	33.57	20.33
	GCE	**92.00**	90.90	**90.80**	89.30	86.70	79.03	91.27	**91.20**	89.73	85.80	78.20	62.07
	JoAPR	90.30	90.45	89.90	88.27	86.93	83.93	90.30	89.30	88.30	88.73	85.80	81.90
	NLPrompt	91.73	**91.13**	90.77	**89.93**	**88.30**	**86.70**	**91.60**	91.17	**90.20**	**89.27**	**86.17**	**81.07**

Table 2. Test accuracy (%) on Food101N.

Method	CoOp	GCE	JoAPR	NLPrompt
Accuracy	69.50	71.32	72.57	76.46

Table 3. The generalization of NLPrompt .

Method/Noise Ratio	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
VPT	89.20	79.43	65.20	61.37	41.67	27.67
VPT+Ours	**91.80**	**91.07**	**89.53**	**86.93**	**80.73**	**73.90**
MaPLe	83.27	77.67	65.27	55.40	37.53	25.47
MaPLe+Ours	**89.23**	**84.30**	**78.37**	**76.43**	**73.30**	**59.87**
PromptSRC	90.61	84.67	78.57	72.27	60.43	49.37
PromptSRC+Ours	**91.29**	**87.67**	**84.97**	**80.33**	**72.10**	**59.50**

ences between the two noise settings and more implementation details are provided in the Supplementary Materials.

6.4. Performance Comparison

For the synthetic noisy datasets, we introduce noise of varying intensities, ranging from 12.5% to 75.0%, with an interval of 12.5%. The experimental results are shown in Table 1. In the vast majority of cases, our NLPrompt achieves state-of-the-art performance, with only a few instances where it performs almost identically to the best results, showing negligible differences. Furthermore, in scenarios with high levels of noise, our method consistently outperforms the other methods, showing a significant performance improvement. This demonstrates the effectiveness and superiority of our method in handling noisy labels in prompt learning. The experimental results on a real-world noisy dataset Food101N are shown in Table 2, where our NLPrompt outperforms all baseline methods, further highlighting the superiority of our approach.

6.5. The Generalization of NLPrompt

Our method is effective not only for CoOp but also for other prompt-tuning approaches, including VPT [27], Maple [29], and PromptSRC [30], which are subsequent methods of CoOp. Additional results on the EuroSAT dataset under symmetric noise are shown in Table 3. It demonstrates the strong generalization ability of NLPrompt.

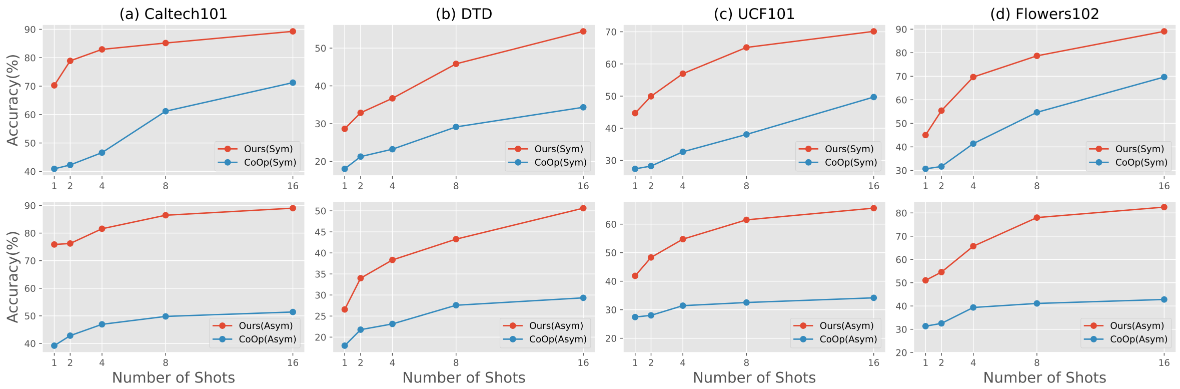

Figure 3. Performance with the different number of shots.

6.6. Few-shot Learning Analysis

Following the few-shot evaluation setting used in previous work [60], we further investigate the impact of the number of shots on different datasets. To this end, we vary the number of shots during training within the range of [1, 2, 4, 8, 16], while keeping the noise rate fixed at 50%. The experimental results are shown in Figure 3, where the horizontal axis represents the number of shots and the vertical axis shows the test accuracy. We observe that as the number of shots increases, the performance of each method improves gradually. However, our method consistently outperforms the others, significantly enhancing the robustness of CoOp across different shot numbers and noise levels.

Table 4. Ablation studies under multiple label noise ratios on Flowers102. (%)

	Method/Noise Ratio	10%	30%	50%	70%	Avg
w\o OT	(a) all data with CE	92.71	86.92	79.36	57.61	79.15
	(b) all data with MAE	88.47	89.07	85.20	80.87	85.90
w\ OT	(c) w\o text feature init	87.16	83.81	79.77	73.00	80.94
	(d) w\o noisy data	84.77	84.53	81.60	77.60	82.16
	(e) w\o clean data	90.17	90.13	88.60	80.55	87.36
	NLPrompt	**96.87**	**93.44**	**92.30**	**85.38**	**92.00**

6.7. Ablation Study

To evaluate the effectiveness of each component of our method, we conduct ablation studies on the Flowers102 dataset. Here, we employ ViT-B/16 as the backbone and train for 100 epochs in the symmetric noise scenario. The experimental results are shown in Table 4. To validate the effectiveness of OT, we designed two sets of experiments: one without using OT for data purification and another using OT for data purification. The experimental design is as follows: (a) Use CE loss for all data; (b) Use MAE loss for all data; (c) Use random initialization prototype instead of CLIP text feature as initialization; (d) Use CE loss for clean data only after removing noisy data; (e) Use MAE loss for noisy data only after removing clean data.

The average results show that (b) outperforms (a), validating the effectiveness of our PromptMAE. Moreover, the average results show that (d) outperforms (a), and (e) outperforms (b), further validating the effectiveness of PromptOT in the data purification process. Additionally, the comparison between (c) and NLPrompt highlights the importance of text feature initialization in our method. Among all methods, our NLPrompt achieves the best performance, with significant improvements over other baselines, further validating the effectiveness of each component.

7. Conclusion

In this study, we addressed the critical challenge of noisy labels in prompt learning for vision-language foundation models by introducing PromptMAE and PromptOT. Our findings demonstrate that adopting the MAE loss in prompt learning—despite its traditionally rare application in noisy-label scenarios—substantially enhances robustness and maintains high accuracy. By leveraging feature learning theory, we elucidated that MAE effectively suppresses the impact of noisy samples, thus improving the overall robustness. Furthermore, the introduction of PromptOT, a prompt-based OT data purification method, allows for an accurate partition of datasets into clean and noisy subsets. This selective application of CE loss to clean data and MAE loss to noisy data in NLPrompt underscores a simple yet powerful strategy for robust prompt learning. Extensive experiments conducted across various noise settings have confirmed the significant performance improvements. NLPrompt capitalizes on the expressive representation and precise alignment capabilities of vision-language models, presenting a promising solution to enhance the robustness of prompt learning in real-world scenarios. Extending NLPrompt to scenarios with unbalanced distributions is under consideration for the future work.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No.62406195, No.62303319), Shanghai Local College Capacity Building Program (23010503100), ShanghaiTech AI4S Initiative SHTAI4S202404, HPC Platform of ShanghaiTech University, MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University) and Shanghai Engineering Research Center of Intelligent Vision and Imaging.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 4

[2] Y. M. Asano, C. Rupprecht, and A. Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020. 4

[3] Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022. 2, 3, 4

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 4

[5] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[6] Wanxing Chang, Ye Shi, Hoang Tuan, and Jingya Wang. Unified optimal transport framework for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35:29512–29524, 2022. 4

[7] Wanxing Chang, Ye Shi, and Jingya Wang. Csot: Curriculum and structure-aware optimal transport for learning with noisy labels. *Advances in Neural Information Processing Systems*, 36:8528–8541, 2023. 4

[8] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[9] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with Instance-Dependent Label Noise: A Sample Sieve Approach. In *International Conference on Learning Representations*, 2021. 2

[10] 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. 6

[11] Tianyu Cui, Hongxia Li, Jingya Wang, and Ye Shi. Harmonizing generalization and personalization in federated prompt learning. In *International Conference on Machine Learning*, 2024. 1

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 3

[13] 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. 6

[14] Chuanwen Feng, Yilong Ren, and Xike Xie. OT-filter: An optimal transport filter for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16164–16174, 2023. 2, 4

[15] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2206–2212, 2021. 2

[16] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9284–9292, 2021. 4

[17] Aritra Ghosh, Himanshu Kumar, and P. Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 1, 2

[18] Dandan Guo, Zhuo Li, He Zhao, Mingyuan Zhou, and Hongyuan Zha. Learning to re-weight examples with optimal transport for imbalanced classification. *Advances in Neural Information Processing Systems*, 35:25517–25530, 2022. 4

[19] 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. 2, 6

[20] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Coteaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018. 2

[21] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian 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*, 12(7):2217–2226, 2019. 6

[22] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019. 2

[23] Wei Huang, Yuan Cao, Haonan Wang, Xin Cao, and Taiji Suzuki. Graph neural networks provably benefit from structural information: A feature learning perspective. *arXiv preprint arXiv:2306.13926*, 2023. 3

[24] Wei Huang, Andi Han, Yongqiang Chen, Yuan Cao, Zhiqiang Xu, and Taiji Suzuki. On the comparison between multi-modal and single-modal contrastive learning.

Advances in Neural Information Processing Systems, 37: 81549–81605, 2024.

[25] Wei Huang, Ye Shi, Zhongyi Cai, and Taiji Suzuki. Understanding convergence and generalization in federated learning through feature learning theory. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[26] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021. 4

[27] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022. 7

[28] Jiarui Jiang, Wei Huang, Miao Zhang, Taiji Suzuki, and Liqiang Nie. Unveil benign overfitting for transformer in vision: Training dynamics, convergence, and generalization. *Advances in Neural Information Processing Systems*, 37: 135464–135625, 2024. 3

[29] 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. 2, 7

[30] 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. 7

[31] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 6

[32] Zhengfeng Lai, Chao Wang, Sen-ching Cheung, and Chen-Nee Chuah. Sar: Self-adaptive refinement on pseudo labels for multiclass-imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4100, 2022. 4

[33] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In *International Conference on Machine Learning*, pages 3763–3772. PMLR, 2019. 2

[34] 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. 6

[35] Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12151–12161, 2024. 1

[36] Junnan Li, Richard Socher, and Steven CH Hoi. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations*, 2020. 2

[37] Yueming Lyu and Ivor W. Tsang. Curriculum loss: Robust learning and generalization against label corruption. *arXiv preprint arXiv:1905.10045*, 2019. 2

[38] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2020. 2

[39] Vu Nguyen, Hisham Husain, Sachin Farfade, and Anton van den Hengel. Confident sinkhorn allocation for pseudo-labeling. *arXiv preprint arXiv:2206.05880*, 2022. 4

[40] 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. 6

[41] Bikang Pan, Wei Huang, and Ye Shi. Federated Learning from Vision-Language Foundation Models: Theoretical Analysis and Method. *arXiv preprint arXiv:2409.19610*, 2024. 2, 3, 4, 5, 6

[42] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012. 6

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[44] Deep Patel and P. S. Sastry. Adaptive sample selection for robust learning under label noise. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3932–3942, 2023. 2

[45] Yuxin Peng, Yunzhen Zhao, and Junchao Zhang. Two-stream collaborative learning with spatial-temporal attention for video classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3):773–786, 2018. 6

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6

[47] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 2

[48] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 4

[49] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pages 5907–5915. PMLR, 2019. 2

[50] Haobo Wang, Mingxuan Xia, Yixuan Li, Yuren Mao, Lei Feng, Gang Chen, and Junbo Zhao. Solar: Sinkhorn label refinery for imbalanced partial-label learning. *Advances in neural information processing systems*, 35:8104–8117, 2022. 4

[51] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021. 3

[52] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021. 4

[53] 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. 1, 2, 6

[54] Jun Xia, Cheng Tan, Lirong Wu, Yongjie Xu, and Stan Z. Li. OT cleaner: Label correction as optimal transport. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3953–3957. IEEE, 2022. 2, 4

[55] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in neural information processing systems*, 32, 2019. 2

[56] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International Conference on Learning Representations*, 2020. 2

[57] Jiangchao Yao, Jiajie Wang, Ivor W. Tsang, Ya Zhang, Jun Sun, Chengqi Zhang, and Rui Zhang. Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, 28(4):1909–1922, 2018. 2

[58] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual T: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems*, 33: 7260–7271, 2020. 2

[59] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5310–5319, 2021. 4

[60] 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. ISBN: 0920-5691 Publisher: Springer. 1, 2, 6, 8

[61] 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. 1, 2

[62] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. 2

[63] Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam in learning neural networks with proper regularization. In *The Eleventh International Conference on Learning Representations*, 2021. 3

NLPrompt: Noise-Label Prompt Learning for Vision-Language Models
Supplementary Materials

Supplementary Organization:

A. Algorithm Framework

NLPrompt employs optimal transport to enhance robust prompt learning by categorizing data into clean and noisy subsets and adapting different loss to each subset. The following pseudo-code illustrates the computation process for NLPrompt:

Algorithm 1 NLPrompt: Optimal Transport-Based Data Partition for Robust Prompt Learning

```
Initialize text encoder $h$, image encoder $\mathbf{g}$, class prompts $\mathbf{p}_c$ and trainable prompt $\mathbf{p}$
for each batch $\{\mathbf{x}_i\}_{i=1}^B$ do
    Compute image features $\mathbf{I}$ and text features $\mathbf{T}$
    Compute similarity matrix $\mathbf{S} = \mathbf{T}\mathbf{I}^\top$
    Solve OT problem (9) to get $\mathbf{Q}^\star$
    Generate pseudo-labels $\tilde{y}_i = \arg\max_j \mathbf{Q}^\star_{ij}$
    Partition data into $\mathcal{D}_{\text{clean}}$ and $\mathcal{D}_{\text{noisy}}$
    for each sample $(\mathbf{x}_i, \tilde{y}_i)$ do
        if $i \in \mathcal{D}_{\text{clean}}$ then
            Use CE loss to update prompts
        else
            Use MAE loss to update prompts
        end if
    end for
end for
return Fine-tuned text prompt $\mathbf{p}$
```

B. Details of Dataset Setup

We selected eight representative visual classification datasets as benchmarks and manually added noise to create synthetic noisy datasets. Additionally, we included a real-world noisy dataset, Food101N, which inherently contains noise and does not require manual modification. Detailed statistics for each dataset, including the original task, the number of classes, and the sizes of training and test samples, are presented in Table A5.

Table A5. The detailed statistics of datasets used in experiments.

Noise Type	Dataset	Task	Classes	Training Size	Testing Size
Synthetic noisy dataset	Caltech101	Object recognition	100	4,128	2,465
	Flowers102	Fine-grained flowers recognition	102	4,093	2,463
	OxfordPets	Fine-grained pets recognition	37	2,944	3,669
	UCF101	Video action recognition	101	7,639	3,783
	DTD	Texture recognition	47	2,820	1,692
	EuroSAT	Satellite image classification	10	13,500	8,100
	StanfordCars	Fine-grained car recognition	196	6,509	8,041
	SUN397	Scene recognition	397	15,880	19,850
Real-world noisy dataset	Food101N	Fine-grained food recognition	101	310,009	30,300

C. Further Experiments

C.1. Experiments on SUN397

We also conducted experiments on SUN397, a dataset with a large number of classes. The results are presented in Table A6. On this dataset, our NLPrompt still outperforms all baseline methods, further highlighting the superiority of our approach.

Table A6. Test accuracy (%) on SUN397.

Dataset	**Method**	**Noise Rate: Sym**						**Noise Rate: Asym**					
		12.5%	**25.0%**	**37.5%**	**50.0%**	**62.5%**	**75.0%**	**12.5%**	**25.0%**	**37.5%**	**50.0%**	**62.5%**	**75.0%**
SUN397	CoOp	65.50	62.90	59.30	55.50	48.30	37.80	63.50	56.10	45.50	33.80	22.10	11.40
	GCE	67.60	66.30	65.40	64.20	62.00	59.20	68.40	66.40	63.80	60.00	53.60	43.80
	NLPrompt	**68.40**	**67.50**	**66.40**	**64.80**	**64.10**	**61.70**	**68.70**	**67.50**	**66.10**	**64.00**	**61.40**	**53.00**

C.2. Experiments on Purfication Strategy

We present experiments to demonstrate that PromptOT is an effective purification method in prompt learning. We frame the purification task as a binary classification problem, where the goal is to distinguish between clean and noisy samples. To evaluate the purification performance, we use accuracy and F1-score metrics for this binary classification task. We compare our purification strategy with the partition method that uses pseudo-labels generated by CLIP zero-shot (denoted as CLIP-ZS), as well as the partition strategy used in JoAPR [19]. For our experiments, we use the Caltech101, DTD, and Flowers datasets, with results shown in Table A7. From this table, we observe that our prompt-based OT selection method achieves higher purification accuracy compared to both the CLIP zero-shot partition and the JoAPR partition.

Table A7. Comparison between different purification strategies. (%)

Method	**Caltech101**		**DTD**		**Flowers**	
	Acc	F1	Acc	F1	Acc	F1
CLIP-ZS	70.38	58.35	56.25	26.72	56.62	25.16
JoAPR	50.88	44.99	51.54	32.32	50.37	32.56
NLPrompt	**75.37**	**67.41**	**59.97**	**36.89**	**62.07**	**39.49**

C.3. Hyperparameter Ablation

We evaluate the impact of different context token lengths on the DTD dataset under 50% symmetric noise in Table A8 and test various entropy regularization coefficients of OT in Figure A4. The results indicate that our NLPrompt is robust to hyperparameters.

Table A8. Test accuracy (%) under different context token lengths.

Context length	1	4	8	16	32
NLPrompt	53.65±1.95	54.00±1.49	54.07±0.40	55.20±0.33	54.73±1.52

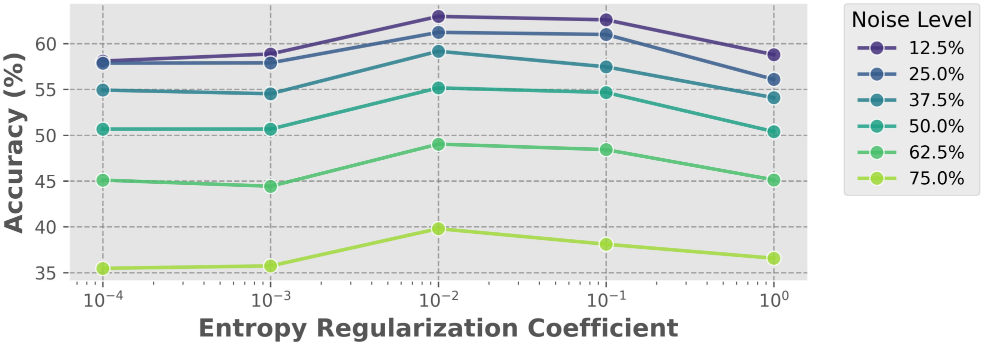

Figure A4. Test accuracy (%) under different entropy regularization coefficients.

C.4. Comparison with JoAPR

The discrepancy in JoAPR results compared to the original paper arises from ***different noise settings***. JoAPR introduces noisy samples uniformly across each class, whereas we randomly distribute noisy samples throughout the entire training set. For example, at a 75% noise rate with 16 shots, JoAPR's setting turns 12 out of 16 samples in each class into noisy samples and ensuring at least 4 clean samples per class. In contrast, our setting may lead to some classes having only 1 or even no clean samples. Thus, our setting is more challenging and practical, leading to significant performance variance at high noise levels. Additionally, we implemented our NLPrompt on the DTD dataset ***using JoAPR's settings*** and consistently observed improvements over JoAPR. The results are shown in Table A9.

Table A9. Comparison with JoAPR in the JoAPR's noise setting.

Method/Noise Ratio	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
JoAPR	58.83	57.67	55.70	53.07	50.67	46.30
NLPrompt	**63.17**	**61.96**	**60.82**	**59.63**	**53.83**	**49.60**

C.5. Comparison with Traditional LNL Method

The traditional LNL method does not fully utilize the benefits of prompt learning with VLMs, whereas our method effectively leverage in prompt learning. In addition to the GCE method discussed in the paper, we also investigate the performance of Mixup on the EuroSAT dataset which is shown in Table A10.

Table A10. Test accuracy (%) compared with Mixup.

Method/Noise Ratio	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%
CoOp	76.50	69.23	61.67	52.33	37.63	26.70
CoOp+Mixup	75.30	71.63	64.07	54.23	42.37	26.90
NLPrompt	**82.53**	**79.53**	**78.13**	**66.70**	**63.53**	**43.80**

C.6. Computational Overhead of Optimal Transport

We evaluate the computation time of optimal transport through experiments. Leveraging the efficient Sinkhorn algorithm for optimal transport, the computational overhead is minimal. For the Caltech101 dataset under a 16-shot learning setup, we perform batch matching, where 1600 image features are matched against 100 classes. The average time required for optimal transport per epoch is 0.00173 seconds, while the original backward process takes 4.352 seconds. Furthermore, even for significantly larger datasets, such as $100,000$ images $\times 1,000$ classes, the average computation time for the optimal transport method is approximately 1.888 seconds, which remains relatively small compared to the overall training time.

D. Limitation

In this paper, we utilize prompt-based optimal transport for data purification, dividing the data into clean and noisy subsets. For the clean data, we apply cross-entropy (CE), while for the noisy data, we employ mean absolute error (MAE). As shown in Table 1, our experimental results indicate that although NLPrompt achieves state-of-the-art performance in most cases, its performance is not always superior to other methods under low noise rates, with slight gaps compared to the best results. This suggests room for improvement in scenarios with low noise levels. This may be because, at low noise levels, optimal transport can misclassify some correct samples as noisy, leading to reduced performance of MAE on datasets with low noise rates.

E. Theoretical Analysis for the Robustness of PromptMAE

In this section, we demonstrate that the mean absolute error (MAE) loss is robust for prompt learning in vision-language foundation models. Leveraging the properties of vision-language pre-trained models, we assume that the latent spaces of the text encoder and image encoder are well-aligned. To clarify, we restate and explain some of our analysis settings. For a classification task, the objective is to classify an image $\mathbf{x}$ into its ground truth class $y \in [C]$, where C represents the total number of classes. For simplicity, we assume that the features corresponding to these classes are orthogonal. In our theoretical analysis, we focus on a binary classification scenario, where $y_i \in \{+1, -1\}$. In most theoretical work of feature learning, it is common to apply insights from binary classification to interpret experimental observations [1, 3]. So, we employ such theoretical frameworks to validate and support our experimental findings.

Outline In our proof, we begin by introducing the assumptions and feature modeling. We then analyze the gradient update using the chain rule and explore the relationship between the feature space and the gradient update. Next, we utilize the decomposition of trainable parameters to demonstrate how the decomposition coefficients change. By establishing the connection between the coefficients and the test loss, we compare the performance of different loss functions by examining the relationship between their respective coefficients.

Following the standard feature learning theory [1], we assume that the weights of the pretrained model consist of two components: task-relevant weights $\boldsymbol{\mu}$ and task-irrelevant weights $\boldsymbol{\xi}$. We begin by proving the following feature representation lemma.

Lemma E.1 (**Restatement of Lemma 4.1: Feature Representation).** *At the t-th iteration, the learnable prompt $\mathbf{p}^{(t)}$ can be rewritten as a linear combination of features and prompt initialization:*

$$\mathbf{p}^{(t)} = \alpha^{(t)}\mathbf{p}^{(0)} + \beta^{(t)}||\boldsymbol{\mu}||_2^{-2}\boldsymbol{\mu} + \sum_{l=1}^{L} \phi_l^{(t)}||\boldsymbol{\xi}_l||_2^{-2}\boldsymbol{\xi}_l,$$

where $\alpha^{(t)}$ are the coefficients of initialization, $\beta^{(t)}$ is the coefficient of task-relevant features, $\phi_{(\cdot)}^{(t)}$ are the coefficients of task-irrelevant features.

Disuccusion on feature decomposition intuition Decomposite the trainable parameters in latent space into linear combinations is a general technique in feature learning theory [1, 3]. Intuitively, in a text prompt, certain "core" words (such as adjectives describing class features before class names) play a key role in determining the image classification and are considered "task-relevant," while other words are "task-irrelevant."

Coefficient dynamics Inspired by the previous study [3], we analyze the dynamics of coefficients in prompt fine-tuning from vision-language foundation models. By analyzing the dynamics of the coefficients, we can reveal the feature learning procedure during training. This analysis allows us to establish the order of coefficients and explore how they are affected by the noisy rate p.

Loss design Our goal here is to compare two different types of loss functions: cross-entropy loss and mean absolute error loss.

$$\ell_{\text{CE}}(\mathbf{s}_i, \mathbf{y}_i) = \sum_{c=1}^{C} -y_{i,c} \log s_{i,c} = -\mathbf{y}_i \log \mathbf{s}_i, \tag{12}$$

$$\ell_{\text{MAE}}(\mathbf{s}_i, \mathbf{y}_i) = \sum_{c=1}^{C} |y_{i,c} - s_{i,c}| = ||\mathbf{y}_\mathbf{i} - \mathbf{s}_\mathbf{i}||_1 \tag{13}$$

Our analysis will be made under the following assumption:

Assumption E.2. Suppose that:

- The number of training samples $N = \Omega(\text{polylog}(d))$, where d is the dimension of learnable prompts.
- The dimension of latent space m is sufficiently large, i.e., $m = \tilde{\Omega}(N)$.
- The learning rate $\eta \leq \tilde{O}(\min\{||\boldsymbol{\mu}||_2^2, \sigma_p^{-2} m^{-1}\})$ and the standard deviation of network weight initialization $\sigma_0 \leq \tilde{O}(mn) \cdot \min\{(||\boldsymbol{\mu}||_2^2, \sigma_p \sqrt{d}^{-1})\}$.
- we assume that $\boldsymbol{\mu}^T \mathbf{p}_{+1} \geq 0 \geq \boldsymbol{\mu}^T \mathbf{p}_{-1}$ which implies a separability condition in the latent space.

Remark. In this assumption, a sufficiently large number of training samples and latent dimension are used to ensure that the network has concentration properties. Meanwhile, a sufficiently small learning rate and appropriate weight initialization are employed to guarantee that gradient descent, in the theoretical analysis, leads to loss convergence.

Gradient analysis. In the definition of the text encoder (3), the incorporation of $\mathbf{W}\mathbf{p}_c$ introduces nonlinearity between the trainable prompt and the class prompt while maintaining the overall function's nonlinear nature. The assumptions regarding the image encoder (5) suggest that task-relevant features differ depending on whether the label is positive or negative, while task-irrelevant features remain arbitrary and independent of the label's polarity. Furthermore, the training loss objective is designed to strengthen the similarity between the image feature $g(\mathbf{x}_i)$ and the text feature generated by the label class prompt $\mathbf{p}_{y_i}$. This section discusses the computational approach used to analyze the performance of the algorithm, focusing on the gradient calculations essential for optimizing the model parameters. Leveraging the properties of the gradient of the Softmax function, we have:

$$\frac{\partial s_{i,c}}{\partial \textbf{sim}(\mathbf{g}_\mathbf{i}, h_c)} = (1 - s_{i,c}) s_{i,c}, \qquad \frac{\partial s_{i,c}}{\partial \textbf{sim}(\mathbf{g}_\mathbf{i}, h_{-c})} = s_{i,-c} s_{i,c}. \tag{14}$$

We can derive the gradient of $s_{i,c}$ with respect to learnable prompt $\mathbf{p}$ with chain rule.

$$\frac{\partial s_{i,c}}{\partial \mathbf{p}} = (1 - s_{i,c}) s_{i,c} \frac{\partial \textbf{sim}(\mathbf{g}_i, \mathbf{h}_c)}{\partial \mathbf{p}} - s_{i,c} s_{i,-c} \frac{\partial \textbf{sim}(\mathbf{g}_i, \mathbf{h}_{-c})}{\partial \mathbf{p}} \tag{15}$$

$$= s_{i,c} s_{i,-c} \left(\frac{\partial \textbf{sim}(\mathbf{g}_i, \mathbf{h}_c)}{\partial \mathbf{p}} - \frac{\partial \textbf{sim}(\mathbf{g}_i, \mathbf{h}_{-c})}{\partial \mathbf{p}} \right). \tag{16}$$

To calculate the gradient, we need the partial derivatives of $\textbf{sim}(\mathbf{g}_i, \mathbf{h}_c)$ with respect to $\mathbf{p}$. Recall from (3) and (6) that the similarity is defined as:

$$\textbf{sim}(\mathbf{g}_i, \mathbf{h}_c) = \langle \sigma(\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_c) - \sigma(-\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_c), g(\mathbf{x}_i) \rangle. \tag{17}$$

The gradient of this similarity can be derived as:

$$\frac{\partial \textbf{sim}(\mathbf{g}_i, \mathbf{h}_c)}{\partial \mathbf{p}} = (\mathbf{W}^T (\sigma'(\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_c) + \sigma'(-\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_c))) \cdot g(\mathbf{x}_i). \tag{18}$$

Using the previously derived gradient, we have

$$\frac{\partial \ell}{\partial \mathbf{p}} = \frac{\partial \ell}{\partial s_{i,\tilde{y}_i}} \frac{\partial s_{i,\tilde{y}_i}}{\partial \mathbf{p}} = \frac{\partial \ell}{\partial s_{i,\tilde{y}_i}} s_{i,\tilde{y}_i} s_{i,-\tilde{y}_i} \left(\frac{\partial \textbf{sim}(\mathbf{g}_i, \mathbf{h}_{\tilde{y}_i})}{\partial \mathbf{p}} - \frac{\partial \textbf{sim}(\mathbf{g}_i, \mathbf{h}_{-\tilde{y}_i})}{\partial \mathbf{p}} \right). \tag{19}$$

The gradient for different loss functions can then be computed as follows:

$$\frac{\partial \ell_{CE}}{\partial s_{i,\tilde{y}_i}} = -\frac{1}{s_{i,\tilde{y}_i}}, \qquad \frac{\partial \ell_{MAE}}{\partial s_{i,\tilde{y}_i}} = -2 \tag{20}$$

Here, we define $\ell'_i = \frac{\partial \ell}{\partial s_{i,\tilde{y}_i}} s_{i,\tilde{y}_i} s_{i,-\tilde{y}_i}$ as the gradient coefficient. For cross-entropy loss, this simplifies to $\ell'_i = s_{i,-\tilde{y}_i}$, while for mean absolute error, it becomes $\ell'_i = 2s_{i,\tilde{y}_i} s_{i,-\tilde{y}_i}$. Defining $\sigma'_{r,i} := \sigma(\mathbf{w}_r^T \mathbf{p} + \mathbf{w}_r^T \mathbf{p}_{\tilde{y}_i}) + \sigma(-\mathbf{w}_r^T \mathbf{p} + \mathbf{w}_r^T \mathbf{p}_{\tilde{y}_i}) - \sigma(\mathbf{w}_r^T \mathbf{p} + \mathbf{w}_r^T \mathbf{p}_{-\tilde{y}_i}) - \sigma(-\mathbf{w}_r^T \mathbf{p} + \mathbf{w}_r^T \mathbf{p}_{-\tilde{y}_i})$, the gradient can be expressed as follows:

$$\nabla_{\mathbf{p}} L_{\mathcal{T}}(\mathbf{p}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{m} \ell'_i x_{r,i} \sigma'_{r,i} \mathbf{w}_r. \tag{21}$$

Due to the update rule of gradient descent and the previous gradient formula, we can rewrite the update equation as follows:

$$\begin{aligned} \mathbf{p}^{(t+1)} &= \mathbf{p}^{(t)} - \eta \nabla_{\mathbf{p}} L_{\mathcal{T}}(\mathbf{p}^{(t)}) \qquad (22) \\ &= \mathbf{p}^{(t)} + \frac{\eta}{n} \sum_{i=1}^{n} \sum_{r=1}^{m} \ell'_i x_{r,i} \sigma'_{r,i} \mathbf{w}_r, \qquad (23) \end{aligned}$$

where $\eta \geq 0$ is the learning rate. Next, we use the assumption about the rows of the weight matrix, as shown in (4). This leads to the update formula for the corresponding rows of the features:

$$\begin{aligned} \beta^{(t+1)} &= \beta^{(t)} + \frac{\eta}{n} \sum_{i=1}^{n} \ell'_i \sigma'_{1,i} \tilde{y}_i y_i ||\boldsymbol{\mu}||_2^2 \qquad (24) \\ &= \beta^{(t)} + \frac{\eta}{n} \sum_{i \in S_+} \ell'_i \sigma'_{1,i} ||\boldsymbol{\mu}||_2^2 - \frac{\eta}{n} \sum_{i \in S_-} \ell'_i \sigma'_{1,i} ||\boldsymbol{\mu}||_2^2 \\ \phi_l^{(t+1)} &= \phi^{(t)} + \frac{\eta}{n} \sum_{i=1}^{n} \ell'_i \sigma'_{1,i} \tilde{y}_i x_{l+1,i} ||\boldsymbol{\xi}_l||_2^2. \qquad (25) \end{aligned}$$

E.1. Theoretical analysis

Our analysis follows this logic: both CE and MAE will increase task-relevant and task-irrelevant features for clean data. However, for noisy data, this leads to a decrease in task-relevant features and an increase in task-irrelevant features, causing the SNR to decrease for both. Inspired by [41], we introduce the following lemma.

Lemma E.3. *Under prompt learning, the test loss can be evaluated with the ratio between the task-relevant coefficient and task-irrelevant coefficient.*

Proof. For simplicity, we first introduce the definitions of F_+ and F_-. Here F_+ means the train loss corresponding to the positive class, while F_- means the train loss corresponding to the negative class.

$$\begin{aligned} F_+(\mathbf{p}) &= \sigma(\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_+) - \sigma(-\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_+), \qquad (26) \\ F_-(\mathbf{p}) &= \sigma(\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_-) - \sigma(-\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_-), \qquad (27) \\ F(\mathbf{p}) &= F_+(\mathbf{p}) - F_-(\mathbf{p}). \qquad (28) \end{aligned}$$

From (28), we have that the following expressions are equivalent:

$$\langle F_{y_i}(\mathbf{p}) - F_{-y_i}(\mathbf{p}), g(\mathbf{x}_i) \rangle \geq 0 \Longleftrightarrow y_i \langle F(\mathbf{p}), g(\mathbf{x}_i) \rangle \geq 0. \tag{29}$$

Note that

$$\mathbf{W}\mathbf{p} = \begin{bmatrix} \boldsymbol{\mu}^\top \mathbf{p} \\ \boldsymbol{\xi}_1^\top \mathbf{p} \\ \vdots \\ \boldsymbol{\xi}_L^\top \mathbf{p} \end{bmatrix} = \begin{bmatrix} \beta ||\boldsymbol{\mu}||_2 \\ \phi_1 ||\boldsymbol{\xi}_1||_2 \\ \vdots \\ \phi_L ||\boldsymbol{\xi}_L||_2 \end{bmatrix}. \tag{30}$$

Also, since the weight and class prompts are fixed during prompt learning, $\mathbf{W}\mathbf{p}_+$ and $\mathbf{W}\mathbf{p}_-$ can be treated as two constant terms. To assess the algorithm's performance, we evaluate the error rate during the testing procedure, which serves as our test loss, denoted as $\ell_{\mathcal{D}}$:

$$\ell_{\mathcal{D}}(\mathbf{p}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(\hat{y}_i = y_i). \tag{31}$$

Recall that the test error is minimized when s_{i,y_i} exceeds $s_{i,-y_i}$. Therefore, the accuracy of the i-th sample is equivalent to:

$$\mathbb{1}(\hat{y}_i = y_i) = \mathbb{P}(s_{i,y_i} - s_{i,-y_i} > 0). \tag{32}$$

According to the monotonicity of the softmax function, we have that

$$s_{i,y_i} - s_{i,-y_i} > 0 \Longleftrightarrow \textbf{sim}(\mathbf{g}_i, h_{y_i}) - \textbf{sim}(\mathbf{g}_i, h_{-y_i}) > 0. \tag{33}$$

Considering each row of $F(\mathbf{p}^{(t)})$ and $\mathbf{x}$, we can expand the expression as:

$$F_1(\mathbf{p}^{(t)}) + \sum_{l=1}^{L} y_i x_{i,l} F_{l+1}(\mathbf{p}^{(t)}) \geq 0, \tag{34}$$

where $x_{i,r}$ are Gaussian random variables with zero mean, and y_i are random variables independent of $x_{i,r}$. As in the test procedure, $F_1(\mathbf{p}), \dots, F_{l+1}(\mathbf{p})$ are fixed. Therefore, based on the definition of prompt learning, the expectation that each sample is correctly classified is determined by the ratio between $F_1(\mathbf{p})$ and the task-irrelevant coefficients $F_2(\mathbf{p}), \dots, F_{L+1}(\mathbf{p})$.

Since $F_1(\mathbf{p})$ is a monotonic function with respect to $\mu^{(t)}$ and $F_{l+1}(\mathbf{p})$ is a monotonic function with respect to $\phi_l^{(t)}$, we can express the following relationship:

$$\ell_D(\mathbf{p}^{(t)}) \sim \frac{F_1(\mathbf{p}^{(t)})}{\sum\limits_{l=1}^{L} F_{l+1}(\mathbf{p}^{(t)})}. \tag{35}$$

Therefore, the test loss can be analyzed by evaluating the ratio between the task-relevant and task-irrelevant coefficients. Besides, due to the assumption that $\boldsymbol{\mu}^T \mathbf{p}_c \geq 0 \geq \boldsymbol{\mu}^T \mathbf{p}_{-c}$ and the random initialization of $\mathbf{Wp}$ such that $\boldsymbol{\mu}^T \mathbf{p} \geq 0$, we conclude that $F_1(\mathbf{p}) \geq 0$. Additionally, the expectation of x_i is zero, and $F_l(\mathbf{p})$ remains constant for any iteration with a fixed $\mathbf{p}$, for all l. As a result, we derive the following inequality:

$$\mathbb{E}(s_{i,y_i} - s_{i,-y_i}) \geq 0 \Longleftrightarrow \mathbb{E}(s_{i,y_i} - (1 - s_{i,y_i})) \geq 0 \Longleftrightarrow \mathbb{E}(s_{i,y_i}) \geq \frac{1}{2}, \tag{36}$$

which can be used to analyze the feature dynamics in prompt learning. □

Based on the previous lemma, we can derive the following theorem.

Theorem E.4 (**Restatement of Theorem 4.2**). *With high probability at test $1 - d^{-1}$, the test loss $\ell_{\mathcal{D}}$ for the prompt trained by MAE is lower than the prompt trained by CE, i.e., $\ell_{\mathcal{D}}(\mathbf{p}_{MAE}) \leq \ell_{\mathcal{D}}(\mathbf{p}_{CE})$. Here, $\mathbf{p}_{MAE}$ and $\mathbf{p}_{CE}$ refer to the text prompt trained using MAE loss and CE loss, respectively.*

Proof. To prove that the MAE achieves better generalization performance, we need to compare the ratio between task-relevant and task-irrelevant coefficients. When the task-relevant features dominate, the algorithm performs better, whereas the dominance of task-irrelevant features leads to worse performance. Based on the iteration formulas in (24) and (25), we can derive the expected updates for β and ϕ:

$$\beta_{\text{CE}}^{(t+1)} = \beta_{\text{CE}}^{(t)} + \eta \left[(1-p)\frac{1}{\mathbb{E}[s_y]} ||\boldsymbol{\mu}||_2^2 - p\frac{1}{1-\mathbb{E}[s_y]} ||\boldsymbol{\mu}||_2^2 \right], \tag{37}$$

$$\phi_{\text{CE}}^{(t+1)} = \phi_{\text{CE}}^{(t)} + \eta \left[(1-p)\frac{1}{\mathbb{E}[s_y]} \sigma_p^2 d + p\frac{1}{1-\mathbb{E}[s_y]} \sigma_p^2 d \right], \tag{38}$$

$$\beta_{\text{MAE}}^{(t+1)} = \beta_{\text{MAE}}^{(t)} + \eta \left[(1-p) \cdot 2 \cdot ||\boldsymbol{\mu}||_2^2 - p \cdot 2 \cdot ||\boldsymbol{\mu}||_2^2 \right], \tag{39}$$

$$\phi_{\text{MAE}}^{(t+1)} = \phi_{\text{MAE}}^{(t)} + \eta \left[(1-p) \cdot 2 \cdot \sigma_p^2 d + p \cdot 2 \cdot \sigma_p^2 d \right]. \tag{40}$$

In this part, we use (34) and the assumption that $\boldsymbol{\mu}^\top \mathbf{p}_{+1} \geq 0 \geq \boldsymbol{\mu}^\top \mathbf{p}_{-1}$ to show that the expectation of $F_1(\mathbf{p}^{(t)})$ is greater than zero. Building on this, we assume that $x_{l,i}$ are Gaussian random variables with zero mean, and y_i are Rademacher random variables independent of $x_{l,i}$. Taking the expectation of the left-hand side of (34), we obtain:

$$\mathbb{E}[F_1(\mathbf{p}^{(t)}) + \sum_{l=1}^{L} y_i x_{i,l} F_{l+1}(\mathbf{p}^{(t)})] = \mathbb{E}[F_1(\mathbf{p}^{(t)})] \geq 0, \tag{41}$$

which simplifies to:

$$\mathbb{E}[s_{i,y_i} - s_{i,-y_i}] \geq 0 \iff \mathbb{E}[s_{i,y_i} - (1 - s_{i,y_i})] \geq 0 \iff \mathbb{E}[2s_{i,y_i} - 1] \geq 0 \iff \mathbb{E}[s_{i,y_i}] \geq \frac{1}{2}. \tag{42}$$

Here, the first inequality follows from the definition of accurate classification, the second inequality comes from the properties of the softmax function, and the fourth inequality arises from the properties of expectation. Based on this, the increment of the coefficients per step can be evaluated using expectation. We get the following expressions for the ratio of updates:

$$\frac{\Delta\beta_{\text{MAE}}^{(t)}}{\Delta\beta_{\text{CE}}^{(t)}} = \frac{(1-p)\frac{1}{\mathbb{E}[s_y]} - p\frac{1}{1-\mathbb{E}[s_y]}}{2-4p} = \frac{1}{2\mathbb{E}[s_y]} \cdot \frac{1 - p\frac{1}{1-\mathbb{E}[s_y]}}{1-2p}, \tag{43}$$

$$\frac{\Delta\phi_{\text{MAE}}^{(t)}}{\Delta\phi_{\text{CE}}^{(t)}} = \frac{(1-p)\frac{1}{\mathbb{E}[s_y]} + p\frac{1}{1-\mathbb{E}[s_y]}}{2s} = \frac{1}{2\mathbb{E}[s_y]} \left(1 - p\frac{2\mathbb{E}[s_y]-1}{1-\mathbb{E}[s_y]}\right). \tag{44}$$

Given that $\mathbb{E}[s_y] > \frac{1}{2}$, we have $\frac{1}{1-\mathbb{E}[s_y]} > 2$ and $2\mathbb{E}[s_y] - 1 > 0$. Additionally, since $0 \leq p \leq \frac{1}{2}$, it follows that:

$$\frac{\Delta\beta_{\text{MAE}}^{(t)}}{\Delta\beta_{\text{CE}}^{(t)}} = \frac{1}{2\mathbb{E}[s_y]} \cdot \frac{1 - p\frac{1}{1-\mathbb{E}[s_y]}}{1-2p} > \frac{1}{2\mathbb{E}[s_y]}, \tag{45}$$

$$\frac{\Delta\phi_{\text{MAE}}^{(t)}}{\Delta\phi_{\text{CE}}^{(t)}} = \frac{1}{2\mathbb{E}[s_y]} \left(1 - p\frac{2\mathbb{E}[s_y]-1}{1-\mathbb{E}[s_y]}\right) < \frac{1}{2\mathbb{E}[s_y]}. \tag{46}$$

Compared to a model trained using cross-entropy, we observe that the task-relevant coefficient of a model trained using mean absolute error (MAE) increases more quickly relative to the task-irrelevant coefficient. We can use induction to prove the following two properties:

$$\frac{\beta_{\text{MAE}}^{(t)}}{\phi_{\text{MAE}}^{(t)}} \geq \frac{\beta_{\text{CE}}^{(t)}}{\phi_{\text{CE}}^{(t)}} \text{ and } \phi_{\text{MAE}}^{(t)} \leq \frac{1}{2\mathbb{E}[s_y]}\phi_{\text{CE}}^{(t)}. \tag{47}$$

Assuming that the induction hypothesis holds at the t-th iteration, and given that the learning rate is sufficiently small, the increase in the task-irrelevant coefficient is also small. Then we can write the following:

$$\frac{\beta_{\text{MAE}}^{(t+1)}}{\phi_{\text{MAE}}^{(t+1)}} = \frac{\beta_{\text{MAE}}^{(t)} + \Delta\beta_{\text{MAE}}^{(t)}}{\phi_{\text{MAE}}^{(t)} + \Delta\phi_{\text{MAE}}^{(t)}} \geq \frac{\beta_{\text{MAE}}^{(t)}}{\phi_{\text{MAE}}^{(t)}} + \frac{\Delta\beta_{\text{MAE}}^{(t)}}{\phi_{\text{MAE}}^{(t)}} \geq \frac{\beta_{\text{CE}}^{(t)}}{\phi_{\text{CE}}^{(t)}} + \frac{\Delta\beta_{\text{CE}}^{(t)}}{2\mathbb{E}[s_y]\phi_{\text{MAE}}^{(t)}} \geq \frac{\beta_{\text{CE}}^{(t)}}{\phi_{\text{CE}}^{(t)}} + \frac{\Delta\beta_{\text{CE}}^{(t)}}{\phi_{\text{CE}}^{(t)}} \geq \frac{\beta_{\text{CE}}^{(t+1)}}{\phi_{\text{CE}}^{(t+1)}}. \tag{48}$$

Additionally, we have:

$$\phi_{\text{MAE}}^{(t+1)} = \phi_{\text{MAE}}^{(t)} + \Delta\phi_{\text{MAE}}^{(t)} < \phi_{\text{MAE}}^{(t)} + \frac{1}{2\mathbb{E}[s_y]}\Delta\phi_{\text{CE}}^{(t)} \leq \frac{1}{2\mathbb{E}[s_y]}\phi_{\text{CE}}^{(t)} + \frac{1}{2\mathbb{E}[s_y]}\Delta\phi_{\text{CE}}^{(t)} = \frac{1}{2\mathbb{E}[s_y]}\phi_{\text{CE}}^{(t+1)}. \tag{49}$$

In conclusion, after a sufficiently large number of iterations, we have:

$$L_{\mathcal{D}}(\mathbf{p}_{\text{CE}}) \geq L_{\mathcal{D}}(\mathbf{p}_{\text{MAE}}), \tag{50}$$

which demonstrates that the test loss for mean absolute error (MAE) in prompt learning with noisy labels is lower than that of cross-entropy loss, highlighting the robustness of MAE. □