Tuning Vision-Language Models with Candidate Labels by Prompt Alignment *

Zhifang Zhang[1], Yuwei Niu[2], Xin Liu[3], and Beibei Li[1] ✉

[1] College of Computer Science, Chongqing University, Chongqing 400044, China
libeibeics@cqu.edu.cn
[2] Hongshen Honor School, Chongqing University, Chongqing 400044, China
[3] School of Computer Science and Engineering, Southeast University, Nanjing 210096, China

Abstract. Vision-language models (VLMs) can learn high-quality representations from a large-scale training dataset of image-text pairs. Prompt learning is a popular approach to fine-tuning VLM to adapt them to downstream tasks. Despite the satisfying performance, a major limitation of prompt learning is the demand for labeled data. In real-world scenarios, we may only obtain candidate labels (where the true label is included) instead of the true labels due to data privacy or sensitivity issues. In this paper, we provide the first study on prompt learning with candidate labels for VLMs. We empirically demonstrate that VLMs can learn from candidate labels through prompt learning. Nonetheless, its performance drops when the label ambiguity increases. In order to improve its robustness, we propose a simple yet effective framework that better leverages the prior knowledge of VLMs to guide the learning process with candidate labels. Specifically, our framework disambiguates candidate labels by aligning the model output with the mixed class posterior jointly predicted by both the learnable and the handcrafted prompt. Besides, our framework can be equipped with various training objectives for learning with candidate labels to further improve their performance. Extensive experiments demonstrate the effectiveness of our proposed framework. Our code is available at https://github.com/zhangzf01/CoOpPLL.

Keywords: Multimodal Models · Prompt Learning · Candidate Labels.

1 Introduction

Vision-language models (VLMs), such as CLIP [20], have become excellent base models in multiple domains, most of which employ a dual-encoder architecture to align the natural images with descriptive texts. Remarkably, this special training pattern has endowed VLMs with superior zero-shot transfer performance on visual recognition tasks. In specific, during the inference, the pre-trained text encoder receives inputs in the form of man-crafted prompts, *e.g.*, *"a photo of*

* This work was supported by Chongqing Science and Technology Bureau (CSTB2022TIAD-KPX0180).

<CLS>.”. Subsequently, all the generated textual embeddings are matched with the visual embedding obtained from the image encoder to predict the image category. However, the powerful zero-shot ability of VLMs was shown to be heavily dependent on the choice of handcrafted prompts, which needs substantial efforts and professional domain knowledge to design [32]. To avoid the manual design of the prompts, *prompt learning* [32] is proposed, which treats the textual prompt as additional learnable parameters and tunes them while keeping all the original parameters of the pre-trained model fixed. Later, the concept of the prompt is extended to visual prompt [9] and multi-modal prompt [10] in VLMs. Overall, there has been increasing attention paid to prompt learning due to its potential to perform significantly better than zero-shot transfer with a few sets of labeled data [32,31,10,9,30].

While prompt learning has demonstrated effectiveness and efficiency in few-shot supervised learning, the true labels must be provided for the training data used in prompt learning. This is a significant defect and will limit the usage of prompt learning in various real-world scenarios because we may be unable to collect accurate labels due to security issues or labeling difficulties. Fortunately, obtaining a set of candidate labels that includes the true label in these situations is easier. As we see, learning with only candidate labels (also widely known as *partial-label learning (PLL)* [23,26,16,5,7,12,13]) is practically significant, which also has arisen in many vital applications such as web mining [15], online annotation [22] and ecoinformatics [14]. Nevertheless, existing PLL methods primarily focus on training a model from scratch, and the effectiveness of PLL in the new training paradigm called prompt learning remains unconfirmed. To bridge this research gap, we, for the first time, explore the validity and potential approaches for prompt learning with candidate labels.

This paper empirically shows that prompt learning combined with the prevailing PLL training objectives in a vanilla way can learn from candidate labels in Figure 1. However, as experimentally suggested, if the candidate labels become more ambiguous, the model's performance will drop significantly. Fortunately, prompt learning is still more robust than linear probe [20], another tuning method that trains a linear classifier on top of a frozen pre-trained model. We conjecture that the reason lies in the prior knowledge brought by the fixed class token, which can keep VLMs from over-fitting to the false-positive labels in the candidate label set and provide VLMs with preferred zero-shot ability, thus mitigating the error accumulation problem [27] in PLL with high label ambiguity.

Therefore, to enhance the robustness of prompt learning with candidate labels, we propose a simple yet effective framework incorporating the handcrafted prompt to distill the model with more comprehensive prior knowledge. Concretely, it dynamically mixes the class posteriors predicted by both the handcrafted and learnable prompt, followed by aligning the mixed class posterior with the model output. Besides, due to the simplicity and flexibility of our framework, it can cooperate with any current PLL training objectives. With our framework, the overall performance of various PLL training objectives has improved by a large margin when tuning VLMs with candidate labels.

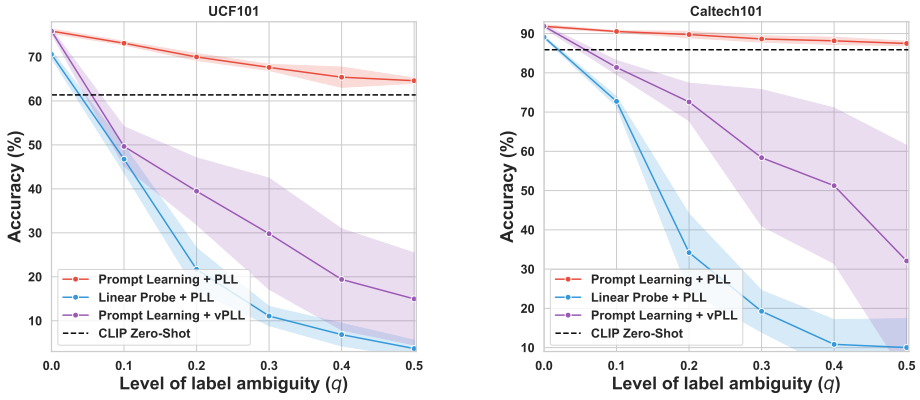

Fig. 1. Performance comparison with multiple fine-tuning approaches combined with PiCO [23] in a vanilla way on UCF101 [21] and Caltech101 [4] with candidate labels of the incremental label ambiguity. vPLL means a simple baseline that treats every candidate label as the ground-truth label and uses cross-entropy loss to learn. We define the level of label ambiguity q as the uniform probability of flipping negative labels $\bar{y}_i \neq y_i$ to false-positive labels inside the candidate label set Y_i: $q = \Pr(\bar{y}_i \in Y_i | \bar{y}_i \neq y_i)$.

Our main contributions can be summarized as follows:

- We provide the first study on the scenario when vision-language models are tuned with only candidate labels.
- We demonstrate empirically and explain that prompt learning combined with PLL training objectives in a vanilla way can learn from candidate labels but is not robust when the label ambiguity is high.
- A framework is proposed to enhance the robustness of prompt learning with candidate labels by aligning the model output with the dynamically mixed prediction by both the handcrafted and the learnable prompt.
- Extensive experiments demonstrate the effectiveness of our framework.

2 Preliminaries of Prompt Learning

CoOp [32] is the first work that migrated prompt learning to vision tasks, which tunes CLIP by optimizing the parameters of the learnable textual prompt (also called soft prompt) while keeping the class token fixed. To be specific, assume CoOp introduces M learnable vectors $\{\boldsymbol{v}_k\}_{k=1}^{M}$ and C fixed class tokens $\{\boldsymbol{c}_l\}_{l=1}^{C}$. Together, they are usually concatenated to form the full prompt $\boldsymbol{s}_i = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_M, \boldsymbol{c}_i\}$ for class i. Let the normalized image embedding be $\boldsymbol{f}^v$, then the class posterior is estimated as:

$$p(y = i|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{f}^v, \text{TextProj}(\boldsymbol{s}_i))/\tau)}{\sum_{j=1}^{C} \exp(\text{sim}(\boldsymbol{f}^v, \text{TextProj}(\boldsymbol{s}_j))/\tau)}. \tag{1}$$

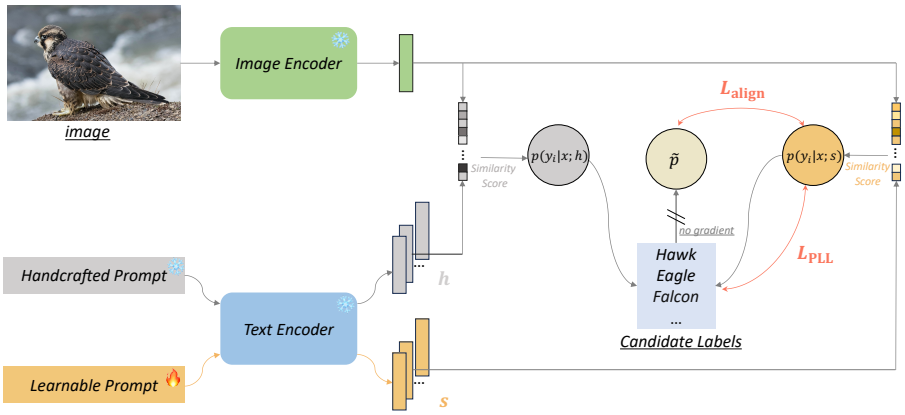

Fig. 2. Illustration of our framework. Our framework includes a prompt alignment module and a PLL method module. All parameters of this framework are **frozen** except for the learnable prompt.

At last, the learnable context vectors $\{\boldsymbol{v}_k\}_{k=1}^{M}$ are optimized on a dataset $D = \{(\boldsymbol{x}_i, y_i)_{i=1}^{N}\}$ with the cross-entropy loss:

$$\mathcal{L}_{\text{true}} = -\mathbb{E}_{(\boldsymbol{x},y)\in D}[\log p(y|\boldsymbol{x})]. \tag{2}$$

Notably, optimizing this training objective of CoOp requires examples of true labels. But in many realistic scenarios, the true label is not accessible. The former research has focused on prompt learning with noisy labels [25] or no label [8,29] and showed prompt learning is not only robust with label noise but also can effectively learn from unlabelled data with specifically designed algorithms.

In this work, we focus on the scenario where only a candidate label set can be obtained, which we unfortunately cannot utilize to optimize the above training objective. Therefore, we will study prompt learning with candidate labels using PLL training objectives instead of cross-entropy loss.

3 Our Framework

This framework is proposed to improve the robustness of prompt learning with candidate labels. It provides a powerful regularization that aligns the dynamically mixed prediction of the handcrafted and learnable prompt with the current model output using weighted cross-entropy loss. It is shown in Figure 2.

Prompt Alignment Regularization.

Let $\mathcal{X}$ be the input space, and $\mathcal{Y} = \{1, 2, \ldots, C\}$ be the label space. The i-th learnable prompt $\boldsymbol{s}_i = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_M, \boldsymbol{c}_i\}$, where $\{\boldsymbol{v}_m\}_{m=1}^{M}$ denotes M learnable tokens and $\boldsymbol{c}_i$ is the the word embedding for the i-th class name. Similarly, the i-th manually crafted prompt is denoted as $\boldsymbol{h}_i$. To clarify, $f_i(\boldsymbol{x}; \boldsymbol{s})$ and $g_i(\boldsymbol{x}; \boldsymbol{h})$ are the softmax outputs of the i-th label, as predicted by the learnable prompt

and handcrafted prompt separately. When $(\boldsymbol{x}, Y)$ is drawn from the partialized dataset, the prompt alignment loss can be calculated as:

$$\mathcal{L}_{\text{align}}(\boldsymbol{x}, Y) = -\sum\nolimits_{i=1}^{C} \tilde{p}_i \log f_i(\boldsymbol{x}; \boldsymbol{s}), \tag{3}$$

where $\tilde{p}_i$ is mixed linearly with the class posteriors predicted by both the learnable and handcrafted prompts:

$$\tilde{p}_i = \alpha p(y = i \mid \boldsymbol{x}; \boldsymbol{s}) + (1 - \alpha) p(y = i \mid \boldsymbol{x}; \boldsymbol{h}). \tag{4}$$

Since the non-candidate labels can never be the ground-truth label, the class posteriors are recalculated as:

$$p(y = i \mid \boldsymbol{x}; \boldsymbol{s}) = \begin{cases} \frac{f_i(\boldsymbol{x}; \boldsymbol{s})}{\sum_{j \in Y} f_j(\boldsymbol{x}; \boldsymbol{s})}, & i \in Y, \\ 0, & i \notin Y. \end{cases} \tag{5}$$

$$p(y = i \mid \boldsymbol{x}; \boldsymbol{h}) = \begin{cases} \frac{g_i(\boldsymbol{x}; \boldsymbol{h})}{\sum_{j \in Y} g_j(\boldsymbol{x}; \boldsymbol{h})}, & i \in Y, \\ 0, & i \notin Y. \end{cases} \tag{6}$$

This regularization can be adapted to any prevailing PLL training objective:

$$\mathcal{L}_{\text{total}}(\boldsymbol{x}, Y) = \mathcal{L}_{\text{PLL}}(\boldsymbol{x}, Y) + \beta \mathcal{L}_{\text{align}}(\boldsymbol{x}, Y), \tag{7}$$

where β is the factor controlling the strength of the alignment loss.

Dynamic Mixing Strategy.

In Equation (4), we use a balancing factor α to mix the handcrafted and learnable prompt predictions. However, a fixed balancing factor may be suboptimal since the prediction quality of the handcrafted prompt will be influenced by several conditions, such as:

- Label Ambiguity: In cases of high label ambiguity, the handcrafted prompt tends to provide more reliable predictions, as soft prompts are difficult to learn effectively in such scenarios.
- Dataset Complexity: For datasets with poor CLIP zero-shot performance, reducing reliance on the handcrafted prompt helps avoid steering the learning process in the wrong direction.
- Training Epoch: If the soft prompt outperforms the handcrafted prompt as training proceeds, it is better to make more use of the soft prompt, ensuring the most effective predictions are leveraged.

Therefore, to further enhance the robustness of our framework against these conditions, a dynamic mixing strategy is adopted to adjust the balancing factor:

$$\tilde{p}_i = \alpha(t) p(y = i \mid \boldsymbol{x}; \boldsymbol{s}) + (1 - \alpha(t)) p(y = i \mid \boldsymbol{x}; \boldsymbol{h}), \tag{8}$$

$$\alpha(t) = \min\{\frac{t}{T'}\lambda, \lambda\}, \tag{9}$$

where T' controls how quickly the balancing factor grows over time, adjusting the shift between the two outputs' contributions and λ sets the maximum weight that the output of soft prompts can have, ensuring its influence is capped. Together, they dynamically balance the contributions of the two prompts for robustness.

4 Experiments

4.1 Experimental Setting

Datasets. We adopt 8 image recognition datasets: ImageNet [3] and Caltech101 [4] for generic object classification; OxfordPets [18], StanfordCars [11], FGVCAircraft [17], and Food101 [1] for fine-grained classification; UCF101 [21] for action classification; DTD [2] for texture classification.

Implementation Details. Our implementation is based on Pytorch [19]. We apply prompt learning on a pre-trained CLIP whose backbone of the image encoder is ResNet-50 [6]. We use the prompt engineering of CLIP [20] to construct the handcrafted prompts. We use a 16-shot fine-tuning strategy and set the learnable prompt token to be 16 with other hyper-parameters the same as CoOp [32]. For the hyperparameters of our method, we set $\lambda = 0.5$, $T' = 25$, $\beta = 1$. Moreover, if a confidence matrix is required in the PLL method, it will be initialized with the model output before training. We report the average test accuracy and the standard deviation of 4 experiments with the seeds fixed.

Training Objectives. We prove the effectiveness of our framework by incorporating six state-of-the-art PLL methods: PRODEN [16], CC [5], LW [24], PiCO [23], PLLCR [26] and CAVL [28].

4.2 Main Results

In Table 1, we compare our framework with vanilla prompt learning for six PLL training objectives on ten benchmark datasets when $q = \{0.1, 0.3, 0.5\}$. For supervised learning, we use the same settings as vanilla prompt learning, except the cross entropy is used, while for CLIP zero-shot, we use the same prompt as the handcrafted prompt in our framework. It is shown that the performance of vanilla prompt learning drops significantly with higher label ambiguity and sometimes even under-performs zero-shot inference, verifying its lack of robustness. On the contrary, our framework not only shows robustness with highly ambiguous candidate labels but also improves general performance that matches supervised learning. For instance, in OxfordPets with LW at $q = 0.5$, vanilla prompt learning under-performs zero-shot inference by 5.84%, while our framework outperforms supervised learning by 2.21%. Moreover, our framework demonstrates consistent improvement across all levels of label ambiguity, not just in cases of high ambiguity. For example, in the Caltech101 dataset, our framework outperforms vanilla prompt learning in 16 out of 18 cases, regardless of the level of label ambiguity and PLL training objectives. Particularly, the more label ambiguity, the more gain in performance with our framework.

The main result has verified the efficacy of our framework. By aligning with the handcrafted prompt, the model can make more accurate predictions at the initial stage of the learning process. Moreover, the class posterior predicted by the handcrafted prompt is hardly affected by the level of label ambiguity, making the learning process more robust.

Table 1. Performance comparison of vanilla prompt learning and our framework. We evaluate the test accuracy of these two methods for different PLL training objectives at different label ambiguities. Supervised means prompt learning with true labels. The standard deviation is shown in parentheses.

	Caltech101			DTD			FGVCAircraft			ImageNet		
Supervised	92.03 (0.22)			62.95 (0.43)			27.27 (3.01)			60.90 (0.64)		
Zero-shot	85.84			42.79			17.07			58.16		
q	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
PRODEN	91.20	88.78	82.05	60.85	54.70	42.15	21.03	9.15	5.58	60.20	47.03	10.18
	(0.37)	(0.11)	(1.37)	(0.42)	(1.35)	(1.23)	(1.65)	(2.11)	(0.97)	(0.19)	(2.26)	(0.19)
+ours	**91.98**	**91.38**	**90.55**	**61.95**	**56.40**	**52.77**	**24.83**	**21.27**	**19.03**	**61.90**	**61.55**	**60.77**
	(0.26)	(0.39)	(0.32)	(1.20)	(0.49)	(2.01)	(1.06)	(0.93)	(0.37)	(0.27)	(0.30)	(0.11)
CC	91.68	91.23	**90.83**	61.10	55.58	48.95	26.00	**22.42**	18.65	61.40	61.08	60.83
	(0.57)	(0.38)	(0.66)	(0.33)	(1.25)	(0.99)	(0.75)	(0.66)	(0.15)	(0.53)	(0.72)	(0.69)
+ours	**92.35**	**91.52**	90.80	**61.25**	**56.25**	**51.77**	**26.00**	22.00	**19.85**	**62.12**	**61.50**	**61.08**
	(0.27)	(0.19)	(0.43)	(0.56)	(0.60)	(0.94)	(0.89)	(0.56)	(0.34)	(0.11)	(0.32)	(0.24)
LW	91.35	89.02	82.97	61.65	54.73	41.77	20.77	9.22	5.85	60.30	45.50	8.80
	(0.26)	(0.37)	(1.13)	(1.11)	(1.84)	(1.36)	(1.15)	(2.04)	(0.93)	(0.82)	(3.52)	(1.58)
+ours	**92.08**	**91.38**	**90.67**	**61.70**	**56.38**	**52.95**	**25.50**	**21.65**	**19.43**	**62.10**	**61.65**	**60.80**
	(0.37)	(0.29)	(0.16)	(0.47)	(0.94)	(1.65)	(0.75)	(0.84)	(0.22)	(0.35)	(0.46)	(0.14)
PLLCR	91.67	91.68	**91.10**	**62.38**	58.25	49.60	24.62	14.60	8.80	60.32	59.60	59.20
	(0.29)	(0.31)	(0.25)	(0.38)	(0.93)	(1.64)	(0.81)	(1.99)	(2.43)	(0.31)	(0.58)	(3.32)
+ours	**91.95**	**91.75**	91.05	62.03	**58.50**	**52.05**	**26.27**	**21.68**	**19.10**	**61.87**	**60.77**	**60.03**
	(0.25)	(0.34)	(0.59)	(1.09)	(0.94)	(0.84)	(0.48)	(0.20)	(0.30)	(0.24)	(0.21)	(0.42)
PiCO	90.47	88.75	87.75	61.70	55.08	48.08	23.60	18.27	14.75	58.98	55.35	52.72
	(0.19)	(0.74)	(0.76)	(1.07)	(1.22)	(1.58)	(2.16)	(2.83)	(3.11)	(0.59)	(0.35)	(0.93)
+ours	**91.92**	**91.33**	**90.72**	**62.67**	**58.42**	**55.08**	**25.12**	**22.18**	**20.32**	**62.03**	**60.92**	**59.80**
	(0.11)	(0.19)	(0.15)	(0.92)	(0.41)	(0.93)	(1.07)	(1.00)	(0.71)	(0.25)	(0.19)	(0.37)
CAVL	91.67	90.97	88.62	**60.85**	50.38	39.80	24.20	7.05	1.88	60.33	58.88	57.55
	(0.40)	(0.51)	(1.30)	(0.97)	(1.45)	(4.97)	(0.51)	(3.50)	(0.79)	0.62	0.94	(1.18)
+ours	**91.82**	**91.58**	**90.73**	60.80	**57.18**	**51.82**	**25.25**	**21.45**	**18.85**	**62.00**	**61.15**	**60.22**
	(0.33)	(0.43)	(0.13)	(0.95)	(0.45)	(1.67)	(0.97)	(0.47)	(0.17)	(0.19)	(0.25)	(0.39)

	OxfordPets			StanfordCars			UCF101			Food101		
Supervised	87.82 (0.43)			71.52 (0.78)			76.55 (0.42)			77.15 (0.17)		
Zero-shot	85.69			55.82			61.88			77.39		
q	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
PRODEN	87.20	86.30	79.08	60.65	12.10	9.53	73.38	63.23	44.83	76.75	76.17	73.58
	(0.45)	(0.70)	(1.52)	(1.03)	(1.88)	(0.66)	(0.77)	(0.86)	(0.76)	(0.54)	(0.49)	(0.47)
+ours	**88.17**	**88.25**	**88.15**	**68.20**	**62.25**	**57.70**	**74.58**	**70.72**	**68.12**	**77.85**	**78.40**	**78.52**
	(0.37)	(0.92)	(0.74)	(0.90)	(1.03)	(1.06)	(0.73)	(0.59)	(0.98)	(0.23)	(0.23)	(0.31)
CC	87.97	87.88	87.45	**70.10**	**66.73**	**62.73**	74.67	71.47	68.95	76.75	76.17	75.45
	(0.62)	(0.31)	(0.27)	(0.60)	(0.26)	(1.16)	(0.26)	(0.35)	(1.46)	(0.54)	(0.49)	(0.92)
+ours	**88.55**	**88.65**	**88.62**	70.03	65.85	62.42	**75.15**	**71.60**	**69.73**	**77.60**	**77.80**	**78.05**
	(0.45)	(0.68)	(0.33)	(0.45)	(0.95)	(0.91)	(0.56)	(0.48)	(0.63)	(0.36)	(0.23)	(0.25)
LW	87.42	86.18	79.85	60.27	12.52	10.85	73.05	62.28	44.75	77.47	76.92	73.05
	(0.41)	(0.91)	(1.20)	(0.28)	(2.11)	(2.16)	(1.28)	(0.68)	(0.69)	(0.44)	(0.61)	(0.73)
+ours	**88.07**	**88.50**	**87.90**	**68.75**	**62.25**	**57.95**	**74.67**	**70.70**	**68.10**	**77.88**	**78.10**	**78.45**
	(0.46)	(0.80)	(0.75)	(0.81)	(0.75)	(0.98)	(0.59)	(0.60)	(0.74)	(0.26)	(0.34)	(0.32)
PLLCR	**86.70**	86.78	85.62	66.25	61.33	25.05	**74.68**	70.22	65.20	76.03	75.70	73.95
	(0.76)	(0.35)	(0.94)	(0.70)	(0.33)	(0.70)	(0.58)	(0.65)	(1.00)	(0.50)	(0.42)	(0.67)
+ours	86.52	**87.27**	**87.27**	**67.05**	**64.55**	**59.85**	74.20	**72.22**	**69.95**	**76.12**	**76.22**	**76.53**
	(0.57)	(0.66)	(0.47)	(0.38)	(0.73)	(1.57)	(0.31)	(0.69)	(1.04)	(0.45)	(0.47)	(0.48)
PiCO	85.90	81.70	78.40	64.27	55.15	49.95	73.45	67.23	64.35	74.03	70.22	67.97
	(1.49)	(1.07)	(1.86)	(0.73)	(0.97)	(1.51)	(0.55)	(0.69)	(0.73)	(0.40)	(1.01)	(1.08)
+ours	**87.65**	**86.78**	**87.15**	**68.55**	**63.67**	**59.55**	**74.80**	**71.65**	**69.55**	**76.65**	**76.05**	**75.68**
	(1.22)	(1.34)	(1.42)	(0.59)	(0.81)	(0.67)	(0.98)	(0.77)	(0.38)	(0.39)	(0.61)	(0.53)
CAVL	87.83	87.88	86.03	69.65	63.75	57.35	73.75	69.30	63.77	76.90	76.75	75.53
	(0.70)	(1.13)	(0.85)	(0.26)	(0.43)	(1.27)	(1.02)	(1.12)	((2.89)	(0.32)	(0.17)	(0.43)
+ours	**88.82**	**88.85**	**88.88**	**69.85**	**65.47**	**61.65**	**74.57**	**71.78**	**69.02**	**77.50**	**78.00**	**78.10**
	(0.48)	(0.27)	(0.25)	(0.61)	(1.15)	(0.78)	(0.63)	(0.48)	(0.78)	(0.16)	(0.16)	(0.23)

4.3 Ablation Studies

In this part, we conduct experiments to assess the effectiveness of our framework with different handcrafted prompts, learnable context lengths, visual backbones, and balancing strategies in 3 datasets: DTD, FGVCAircraft and Caltech101.

Impact of Different Handcrafted Prompts. Because our framework incorporates handcrafted prompts, it is crucial to determine the performance of our framework with differently crafted prompts. In Table 2, we design some prompts with zero-shot performance lower than the default handcrafted prompts [20] and evaluate our framework with these prompts at different levels of label ambiguity. *Introducing a handcrafted prompt can consistently improve the performance, regardless of the quality of the prompt.* Better handcrafted prompts have resulted in better test accuracy in our experiments, which indicates that the handcrafted prompts indeed guide the fine-tuning process with candidate labels. Nonetheless, the performance is improved by a large margin even when we simply set the handcrafted prompt "<CLS>.".

Table 2. Performance comparison of our framework for different handcrafted prompts.

Dataset	Handcrafted Prompt	q			CLIP-Zeroshot
		0.1	0.3	0.5	
DTD	without handcrafted prompt (only prompt learning)	61.70±1.07	55.08±1.22	48.08±1.58	-
	"a photo of a <CLS>."	62.15±0.63	56.85±0.63	53.20±0.91	40.36
	"<CLS>."	62.45±0.18	56.67±0.98	53.23±1.20	41.13
	"<CLS> texture."	**62.67±0.92**	**58.42±0.41**	**55.08±0.93**	**42.79**
FGVC Aircraft	without handcrafted prompt (only prompt learning)	23.60±2.16	18.27±2.83	14.75±3.11	-
	"a photo of a <CLS>."	24.70±1.16	21.82±1.42	19.60±0.81	15.84
	"<CLS>."	24.65±0.89	20.85±0.84	19.02±0.71	15.54
	"a photo of a <CLS>, a type of aircraft."	**25.12±1.07**	**22.18±1.00**	**20.32±0.71**	**17.07**
Caltech101	without handcrafted prompt (only prompt learning)	90.47±0.19	88.75±0.74	87.75±0.76	-
	"<CLS>."	91.42±0.19	90.97±0.51	90.05±0.73	81.34
	"a photo of a <CLS>."	**91.92±0.11**	**91.33±0.19**	**90.72±0.15**	**85.84**

Effectiveness of the Dynamic Mixing Strategy. This experiment is conducted to evaluate the efficacy of our dynamic and mixing strategy. $\alpha(t)$ is set using the same hyperparameters as the main experiment. The result is shown in Table 3. *The dynamic mixing strategy optimizes the model's performance.* The model with dynamic technique outperforms all others in 6 and ranked second in 2, out of 9 cases, demonstrating its effectiveness. When only one handcrafted or learnable prompt is leveraged, the model's performance declines and cannot surpass the performance achieved when $\alpha = 0.7$. This validates the rationality of the mixing strategy. Lastly, when $\alpha = 0$, it outperforms models with only learnable prompt guidance ($\alpha = 1$) or no guidance (vanilla prompt learning) in most

Table 3. Performance comparison of our framework for different α.

	DTD			FGVCAircraft			Caltech101		
q	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
$\alpha = 0$	62.05±1.17	57.12±0.98	53.75±1.41	24.68±0.83	22.10±0.56	19.93±0.31	91.70±0.23	91.20±0.16	**90.93±0.22**
$\alpha = 1$	61.83±0.40	56.45±2.29	50.98±3.08	25.15±0.98	18.83±4.24	15.35±3.76	91.62±0.25	89.83±0.68	88.47±0.37
$\alpha = 0.7$	62.28±0.33	**59.38±1.43**	54.57±1.14	**25.77±0.99**	22.02±0.98	19.95±0.38	91.85±0.23	91.03±0.54	90.35±0.63
$\alpha(t)$	**62.67±0.92**	58.42±0.41	**55.08±0.93**	25.12±1.07	**22.18±1.00**	**20.32±0.71**	**91.92±0.11**	**91.33±0.19**	90.72±0.15

cases, which, to some extent, justifies the effectiveness of explicitly adopting the handcrafted prompt to guide the learning process.

5 Conclusion

This work, for the first time, investigated the scenario when tuning vision-language models (VLMs) with candidate labels. Throughout a series of experiments, we empirically demonstrated that prompt learning combined with PLL training objectives in a vanilla way can learn from candidate labels. However, as the ambiguity of candidate labels increases, its performance is degraded. To alleviate this issue, we proposed a framework that enhances the robustness by aligning the dynamically mixed class posterior of the handcrafted and learnable prompt with the model's output to guide the learning process with candidate labels. Comprehensive experimental results and analysis on multiple benchmark datasets demonstrate the effectiveness of our proposed framework.

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: ECCV. pp. 446–461. Springer (2014)
2. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: CVPR. pp. 3606–3613 (2014)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
4. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: CVPR. pp. 178–178. IEEE (2004)
5. Feng, L., Lv, J., Han, B., Xu, M., Niu, G., Geng, X., An, B., Sugiyama, M.: Provably consistent partial-label learning. NeurIPS pp. 10948–10960 (2020)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
7. Hu, J., Shu, S., Li, B., Xiang, T., He, Z.: An unbiased risk estimator for partial label learning with augmented classes. ACM Trans. Intell. Syst. Technol. **15**(6) (2024)
8. Huang, T., Chu, J., Wei, F.: Unsupervised prompt learning for vision-language models. arXiv preprint arXiv:2204.03649 (2022)
9. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: ECCV. pp. 709–727. Springer (2022)

10. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: CVPR. pp. 19113–19122 (2023)
11. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: ICCV Workshops. pp. 554–561 (2013)
12. Li, B., et al: Asyco: An asymmetric dual-task co-training model for partial-label learning (2024), https://arxiv.org/abs/2407.15036
13. Li, B., et al.: Gemini: A dual-task co-training model for partial label learning. In: AI 2023: Advances in Artificial Intelligence. Springer Nature Singapore (2024)
14. Liu, L., Dietterich, T.: A conditional multinomial mixture model for superset label learning. NeurIPS **25** (2012)
15. Luo, J., Orabona, F.: Learning from candidate labeling sets. NeurIPS **23** (2010)
16. Lv, J., Xu, M., Feng, L., Niu, G., Geng, X., Sugiyama, M.: Progressive identification of true labels for partial-label learning. In: ICML. pp. 6500–6510 (2020)
17. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
18. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: CVPR. pp. 3498–3505. IEEE (2012)
19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library (2019)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
21. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
22. Tang, C.Z., Zhang, M.L.: Confidence-rated discriminative partial label learning. In: AAAI (2017)
23. Wang, H., Xiao, R., Li, Y., Feng, L., Niu, G., Chen, G., Zhao, J.: Pico: Contrastive label disambiguation for partial label learning. In: ICLR (2021)
24. Wen, H., Cui, J., Hang, H., Liu, J., Wang, Y., Lin, Z.: Leveraged weighted loss for partial label learning. In: ICML. pp. 11091–11100. PMLR (2021)
25. Wu, C.E., et al: Why is prompt tuning for vision-language models robust to noisy labels? In: ICCV. pp. 15488–15497 (2023)
26. Wu, D.D., Wang, D.B., Zhang, M.L.: Revisiting consistency regularization for deep partial label learning. In: ICML. pp. 24212–24225. PMLR (2022)
27. Yao, Y., Gong, C., Deng, J., Yang, J.: Network cooperation with progressive disambiguation for partial label learning. In: ECML PKDD. Springer (2021)
28. Zhang, F., Feng, L., Han, B., Liu, T., Niu, G., Qin, T., Sugiyama, M.: Exploiting class activation value for partial-label learning. In: ICLR (2021)
29. Zhang, J., Wei, Q., Liu, F., Feng, L.: Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data. In: ICML (2024)
30. Zhang, Z., He, S., Shen, B., Feng, L.: Defending multimodal backdoored models by repulsive visual prompt tuning. arXiv preprint arXiv:2412.20392 (2024)
31. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)
32. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (2022)