Rethinking Open-world Prompt Tuning: A Systematic Framework for Evaluation and Optimization

Mengwei Li[1], **Zilei Wang**[1*], **Yixin Zhang**[2]

[1]University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
curry@mail.ustc.edu.cn, zlwang@ustc.edu.cn, zhyx12@ustc.edu.cn

Abstract

Prompt Tuning (PT) is a widely used strategy for adapting pre-trained Vision-Language Models (VLMs) to various downstream tasks. Conventional PT methods evaluate performance separately on known (base) and unknown (new) classes. However, in real-world scenarios, models often encounter inputs without prior knowledge of their class domain. This challenge has motivated the development of Open-world Prompt Tuning (OPT), which requires models to first determine whether a sample belongs to base or new classes and then classify it accordingly. In this work, we carefully review existing OPT methods and identify three key limitations: **(L1)** incomplete evaluation metrics, **(L2)** time-consuming and memory-intensive OOD detection methods, and **(L3)** insufficiently comprehensive optimization strategies. To address these issues, we first tackle **L1** by proposing two novel metrics to explicitly evaluate adaptability and generalization under the OPT setting, forming a more comprehensive evaluation framework. For **L2**, we propose a training-free OOD detection method called Entropy-weighted Rank-normalized Fusion (ERF), which first applies rank normalization to both the maximum and the sum of base-class probabilities, followed by an entropy-weighted fusion of the normalized values. For **L3**, we propose a plug-and-play Gated Dual-Merging (GDM) strategy to strengthen the classifier's capability. GDM performs selective merging at the weight level based on an adaptive criterion and combines fine-tuned and LLM-boosted logits at the output level. Extensive experiments on three PT baselines across 11 datasets demonstrate the effectiveness of our proposed ERF and GDM.

Introduction

Vision-Language Models (VLMs), such as CLIP (Radford et al. 2021), have demonstrated strong generalization and transferability on a variety of downstream tasks (Singh et al. 2022; Xiao, Wang, and Li 2024). Although VLMs demonstrate strong capabilities in recognizing open-set visual concepts, their generalization performance degrades significantly when there are substantial class or domain shifts between the upstream training data and downstream tasks.

Motivated by the success of prompt engineering in natural language processing, Prompt Tuning (Zhou et al. 2022b) (PT) has emerged as an effective paradigm for adapting VLMs to open-world scenarios using only a small set of learnable textual or visual prompt vectors. A common setting in PT is base-to-new generalization, where the model is trained on labeled data from base classes and evaluated separately on both base and new classes. However, this setting assumes prior knowledge of class domains (base or new), which limits its practicality in real-world scenarios where such information is typically unavailable. To address this limitation, DeCoOp (Zhou et al. 2024) introduces a more practical setting called Open-world Prompt Tuning (OPT), which evaluates model performance on a mixture of base and new classes while training only on base classes. The task is divided into two stages: the first identifies whether an input belongs to a base or new class, and the second classifies it into the correct category. Building on this, GMoP (Hua et al. 2025) argues that existing evaluation metrics for the OPT task are inadequate, and proposes a new metric, OpenworldAUC, which provides a unified evaluation of base-to-new detection, base class classification, and new class classification.

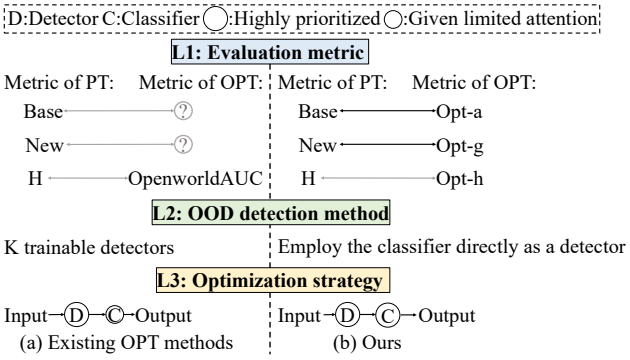

Figure 1: Comparison diagram of our approach and existing open-world prompt tuning methods.

In this paper, we revisit the OPT task and identify three key limitations in existing approaches (depicted in Figure 1): **(L1)** incomplete evaluation metrics, **(L2)** time-consuming and memory-intensive OOD detection methods, and **(L3)** insufficiently comprehensive optimization strategies. For **L1**, we point out that the existing evaluation metric for the OPT

*Corresponding author.

task, OpenworldAUC, only reflects overall performance and lacks metrics that explicitly assess adaptability and generalization, as in PT tasks. To address this, we propose two new metrics specifically designed to evaluate these capabilities under the OPT setting. These metrics complement OpenworldAUC and together form a more comprehensive evaluation framework for the OPT task. For **L2**, existing OPT methods, including DeCoOp and GMoP, typically partition base classes into simulated base and new classes and adopt a meta-learning-like strategy to train an OOD detector. To enhance robustness, they perform k different class splits and train one detector per split. As each sample must be accessed $k+1$ times per epoch, the training becomes time-consuming. Moreover, jointly optimizing k detectors and a classifier significantly increases memory consumption, particularly on datasets with a large number of categories, such as SUN397 and ImageNet. In this work, we propose a training-free OOD detection method called Entropy-weighted Rank-normalized Fusion (ERF), which first applies rank normalization to both the maximum and the sum of base-class probabilities, followed by an entropy-weighted fusion of the normalized values. ERF is computationally efficient and outperforms existing training-based approaches. For **L3**, existing OPT methods primarily focus on enhancing OOD detection to improve overall performance, often overlooking the role of the classifier itself. As shown in Eq. 6, strengthening classification performance can also significantly improve its overall performance. To bridge this gap, we propose a plug-and-play Gated Dual-Merging (GDM) approach. Specifically, at the weight level, we dynamically decide whether to merge the fine-tuned model with the pre-trained model based on our proposed adaptive criterion ΔD; at the logits level, we combine the fine-tuned logits with the LLM-boosted logits. Experimental results demonstrate that this strategy significantly enhances the classification performance of existing PT methods on both base and new classes, thereby improving overall model effectiveness. Our main contributions are as follows:

- We propose two novel metrics to evaluate model adaptability and generalization in the OPT task, thereby further improving its evaluation framework.
- We propose a new training-free OOD detection method based on entropy-guided dynamic weighting of rank-normalized probabilities.
- We introduce a plug-and-play gated dual-merging approach that significantly enhances the classification ability of existing PT methods.
- Extensive experiments on three PT baselines across 11 datasets demonstrate the effectiveness of our approach.

Preliminaries

Prompt Tuning

Prompt Tuning (PT) (Zhou et al. 2022b) is a parameter-efficient paradigm for adapting Vision-Language Models (VLMs) to downstream tasks by optimizing task-specific prompts using limited training data from base classes, while keeping the VLM weights fixed. Let $\boldsymbol{f}_v$ denote the visual feature of the input image $\boldsymbol{x}$ extracted by the image encoder. To obtain class prototypes, the i-th class token is concatenated with learnable prompts to construct the class prompt. The corresponding text feature $\boldsymbol{f}_{t_i}$ is then extracted by feeding this class prompt into the text encoders. The image-class matching score is computed using the cosine similarity $Sim(\boldsymbol{f}_v, \boldsymbol{f}_{t_i})$, and the probability of the i-th class is calculated as:

$$p(\boldsymbol{y}=i|\boldsymbol{x}) = \frac{\texttt{exp}(Sim(\boldsymbol{f}_v, \boldsymbol{f}_{t_i})/\tau)}{\sum_{j=1}^{C} \texttt{exp}(Sim(\boldsymbol{f}_v, \boldsymbol{f}_{t_j})/\tau)}, \quad (1)$$

where τ is the temperature determined by VLMs and C denotes the total number of classes. Finally, the learnable parameters are optimized using the cross-entropy loss $\mathcal{L}_{ce} = -\sum_i \boldsymbol{y}_i \log p(\boldsymbol{y}=i|\boldsymbol{x})$, where $\boldsymbol{y}_i = 1$ if the image x belongs to class i, and $\boldsymbol{y}_i = 0$ otherwise.

After training, the model is evaluated separately on the base dataset $\mathcal{S}_b = \{(\boldsymbol{x}_b^{(i)}, y_b^{(i)})\}_{i=1}^{N_b}$ and new dataset $\mathcal{S}_n = \{(\boldsymbol{x}_n^{(i)}, y_n^{(i)})\}_{i=1}^{N_n}$. Let $\boldsymbol{s}_b \in \mathbb{R}^{C_b}$ and $\boldsymbol{s}_n \in \mathbb{R}^{C_n}$ are the logit scores of the base and new dataset, where C_b and C_n are the numbers of classes in each dataset. The classification accuracy on $\mathcal{S}_b$ and $\mathcal{S}_n$ can be computed as follows:

$$\begin{aligned} \text{Base} &:= \frac{1}{N_b} \sum_{(\boldsymbol{x}_b, y_b)\in\mathcal{S}_b} \mathbf{1}[y_b = \arg\max_{y'\in\mathcal{Y}_b} \boldsymbol{s}_b], \\ \text{New} &:= \frac{1}{N_n} \sum_{(\boldsymbol{x}_n, y_n)\in\mathcal{S}_n} \mathbf{1}[y_n = \arg\max_{y'\in\mathcal{Y}_n} \boldsymbol{s}_n], \end{aligned} \quad (2)$$

where $\mathcal{Y}_b$ and $\mathcal{Y}_n$ denote the label spaces of base and new classes, respectively, and $\mathbf{1}[\cdot]$ is the indicator function. Then, H is defined by the harmonic mean of Base and New:

$$\text{H} := \frac{2 \times \text{Base} \times \text{New}}{\text{Base} + \text{New}}. \quad (3)$$

Open-world Prompt Tuning

Unlike traditional PT, Open-world Prompt Tuning (OPT) evaluates model performance on a mixture of base and new classes. To tackle this challenge, DeCoOp (Zhou et al. 2024) employs a divide-and-conquer strategy, decomposing the task into two stages: base-to-new detection and domain-specific classification. In the first stage **(S1)**, a base-to-new detector r determines whether an input sample belongs to the base or new domain based on a threshold t, where a higher $r(\boldsymbol{x})$ indicates greater likelihood of the sample originating from the base domain. The second stage **(S2)** involves base-class discriminability and new-class discriminability. For samples classified as belonging to the base domain by the OOD detector, the model employs a classifier g based on fine-tuned textual features. In contrast, for samples identified as the new domain, it uses a classifier h based on zero-shot textual features, as fine-tuned models often overfit to the base domain and underperform on new class classification.

Recently, GMoP (Hua et al. 2025) pointed out that existing evaluation metrics—H, OverallAcc, and AUROC (Yang et al. 2021)—do not simultaneously satisfy the three essential requirements of the OPT task: comprehensive assessment of first-stage detection, second-stage classification,

and robustness to domain distribution. To address this, they proposed a new metric, OpenworldAUC, which effectively meets all three criteria. Specifically, they first define the Miss Rate of the base classes and the Hit Rate of the new classes as follows:

$$\begin{aligned}\text{MissRate}_b :=& \mathbb{E}_{\mathcal{S}_b}[\mathbf{1}[r(\boldsymbol{x}_b) \leq t] + \mathbf{1}[r(\boldsymbol{x}_b) > t, y_b \neq g(\boldsymbol{x}_b)]], \\ \text{HitRate}_n :=& \mathbb{E}_{\mathcal{S}_n}[\mathbf{1}[r(\boldsymbol{x}_n) \leq t, y_n = h(\boldsymbol{x}_n)]].\end{aligned} \tag{4}$$

They then define the area under the MissRate_b–HitRate_n curve as a new metric for the OPT task:

$$\text{OpenworldAUC} := \int_{t=0}^{t=1} \text{HitRate}_n(t) \cdot d\text{MissRate}_b(t). \tag{5}$$

Finally, they demonstrated that OpenworldAUC equals to the joint probability that r ranks a base sample $\boldsymbol{x}_b$ above a new sample $\boldsymbol{x}_n$, while classifiers g and h correctly predict $\boldsymbol{x}_b$ and $\boldsymbol{x}_n$, respectively.

$$\mathbb{E}_{\mathcal{D}}\big[\underbrace{\mathbf{1}[y_b = g(\boldsymbol{x}_b)]}_{\textbf{Base (S2\_b)}} \cdot \underbrace{\mathbf{1}[r(\boldsymbol{x}_b) > r(\boldsymbol{x}_n)]}_{\textbf{Base-to-New (S1)}} \cdot \underbrace{\mathbf{1}[y_n = h(\boldsymbol{x}_n)]}_{\textbf{New (S2\_n)}}\big]. \tag{6}$$

Method

Comprehensive Evaluation

A closer look at the evaluation metrics used in the PT task and OPT task reveals that the OPT metric OpenworldAUC corresponds to the PT metric H, as both reflect the overall performance of the model. However, the OPT task lacks counterparts to the Base and New metrics in the PT task, which explicitly evaluate the model's adaptability and generalization ability. This motivates us to propose two new metrics to assess these abilities under the OPT setting. Inspired by OSCR (Dhamija, Günther, and Boult 2018), OpenAUC (Wang et al. 2022) and GMoP (Hua et al. 2025), we first assign all base/new classes to one base/new super-class, respectively. Then we define the Miss Rate of the new super-class and the Hit Rate of the base classes as follows:

$$\begin{aligned}\text{MissRate}_{n-s} :=& \mathbb{E}_{\mathcal{S}_n}[\mathbf{1}[r(\boldsymbol{x}_n) > t]], \\ \text{HitRate}_b :=& \mathbb{E}_{\mathcal{S}_b}[\mathbf{1}[r(\boldsymbol{x}_b) > t, y_b = g(\boldsymbol{x}_b)]].\end{aligned} \tag{7}$$

We treat the area under the curve as a metric to evaluate the model's adaptation ability:

$$\text{Opt-a} := \int_{t=0}^{t=1} \text{HitRate}_b(t) \cdot d\text{MissRate}_{n-s}(t). \tag{8}$$

Similarly, we define the Miss Rate of the base super-class and the Hit Rate of the new classes as follows:

$$\begin{aligned}\text{MissRate}_{b-s} :=& \mathbb{E}_{\mathcal{S}_b}[\mathbf{1}[r(\boldsymbol{x}_b) \leq t]], \\ \text{HitRate}_n :=& \mathbb{E}_{\mathcal{S}_n}[\mathbf{1}[r(\boldsymbol{x}_n) \leq t, y_n = h(\boldsymbol{x}_n)]].\end{aligned} \tag{9}$$

We treat the area under the curve as a metric to evaluate the model's generalization ability:

$$\text{Opt-g} := \int_{t=0}^{t=1} \text{HitRate}_n(t) \cdot d\text{MissRate}_{b-s}(t). \tag{10}$$

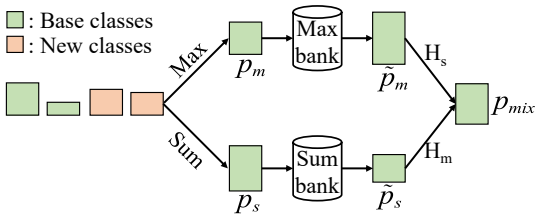

Figure 2: A schematic diagram of ERF. The max bank and sum bank represent the sets of maximum probabilities and summed probabilities across all test samples, respectively.

Finally, substituting the results from the above formulas into the following equation yields the overall performance metric:

$$\text{Opt-h} := \int_{t=0}^{t=1} \text{HitRate}_n(t) \cdot d(1 - \text{HitRate}_b(t)). \tag{11}$$

It is straightforward to prove that Opt-h is equivalent to OpenworldAUC. In summary, the two metrics we propose, Opt-a and Opt-g, complement OpenworldAUC and together form a comprehensive evaluation framework for OPT task.

Entropy-weighted Rank-normalized Fusion

Existing OPT methods generally split base classes into simulated base and new classes, then utilize a meta-learning-like strategy to train an OOD detector. Our experiments show that this approach is both time-consuming and memory-intensive. We propose Entropy-weighted Rank-normalized Fusion (ERF), a training-free and computationally efficient OOD detection method that outperforms existing training-based approaches. Inspired by NegLabel (Jiang et al. 2024), we treat the new classes as negative labels and compute the predicted probabilities of image $\boldsymbol{x}$ across all classes, denoted as $\boldsymbol{p} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_C\}$. We then obtain two naive OOD scores: the maximum probability among base classes (denoted as $\boldsymbol{p}_m$) and the sum of all base class probabilities (denoted as $\boldsymbol{p}_s$). When applied to OOD detection, the performance of these two scores varies across datasets: in some cases, the first outperforms the second, while in others, the second shows superior performance (as shown in Figure 4). Simply mixing up (Zhang et al. 2017) the two scores fails to yield satisfactory results. Therefore, we aim to develop a method that adaptively balances the two scores to achieve consistently strong performance across different datasets.

Initially, we observed that $\boldsymbol{p}_s$ and $\boldsymbol{p}_m$ have different scales; specifically, $\boldsymbol{p}_s$ is always greater than or equal to $\boldsymbol{p}_m$, which may interfere with the mixup operation. Therefore, we apply rank normalization to these two scores. Specifically, we first collect all $\boldsymbol{p}_s$ and $\boldsymbol{p}_m$ values from the test data and sort them in ascending order, respectively. Each value is then normalized by computing the ratio of its index in the sorted list to the total number of values. The normalization is defined by the following formula:

$$\tilde{\boldsymbol{p}}^i = \frac{rank\,(i)}{N}, rank\,(i) = \sum_{j=1}^{N} \mathbf{1}\left(\boldsymbol{p}^j \leqslant \boldsymbol{p}^i\right), \tag{12}$$

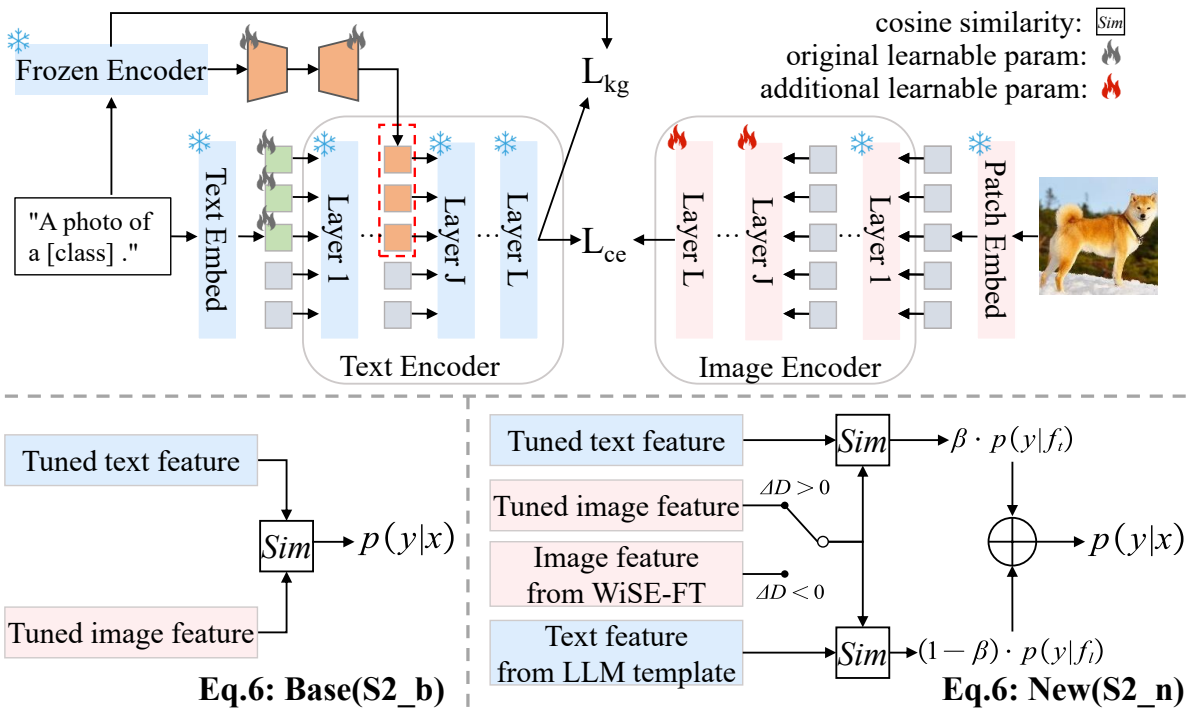

Figure 3: An overview of our proposed GDM framework. Taking the PT method TCP as an example, we first follow its original procedure to obtain a preliminary fine-tuned model. Next, we unfreeze the last six layers of the visual encoder while freezing all other parameters. These unfrozen layers are then optimized using the cross-entropy loss. We use the fine-tuned model to classify base classes directly, while new classes are classified using our gated dual-merging.

where $N = N_b + N_n$ denotes the total number of test samples. Using the above formula, we obtain the normalized $\tilde{\boldsymbol{p}}_s$ and $\tilde{\boldsymbol{p}}_m$. Subsequently, by analyzing the computation of $\boldsymbol{p}_s$ and $\boldsymbol{p}_m$, we find that $\boldsymbol{p}_s$ represents the probability that a sample belongs to any base class, corresponding to a binary classification task (base super-class and new super-class). In contrast, $\boldsymbol{p}_m$ denotes the probability that the sample belongs to a specific base class, aligning with a C-class classification task. The entropies for the binary and C-class classification can be calculated using the following formulas:

$$\mathrm{H}_s = \frac{-\boldsymbol{p}_s \log \boldsymbol{p}_s - (1-\boldsymbol{p}_s)\log(1-\boldsymbol{p}_s)}{\log 2}, \mathrm{H}_m = \frac{-\sum_i \boldsymbol{p}_i \log \boldsymbol{p}_i}{\log C}. \quad (13)$$

Entropy can partially reflect the confidence of a model's predictions: lower entropy indicates higher confidence. To integrate the two predictions $\tilde{\boldsymbol{p}}_s$ and $\tilde{\boldsymbol{p}}_m$, we adopt an entropy-based weighting approach that assigns higher weights to predictions with lower entropy (*i.e.*, higher confidence). The weights are defined as the inverse of the corresponding entropies, normalized, and used to compute a weighted sum for the final prediction. Finally, the simplified entropy-based weighting formula can be expressed as follows:

$$\boldsymbol{p}_{mix} = \frac{\mathrm{H}_m}{\mathrm{H}_s + \mathrm{H}_m} \cdot \tilde{\boldsymbol{p}}_s + \frac{\mathrm{H}_s}{\mathrm{H}_s + \mathrm{H}_m} \cdot \tilde{\boldsymbol{p}}_m. \quad (14)$$

Gated Dual-Merging

Existing OPT approaches, including DeCoOp and GMoP, focus solely on enhancing OOD detection to improve model performance Opt-h. However, as shown in Eq. 6, improving classification accuracy on both base and new classes can also effectively boost overall performance. This motivates us to explore an approach that can effectively enhance the classification capability of existing methods. WiSE-FT (Wortsman et al. 2022) is a simple yet effective method for improving the performance of fine-tuned models. While it has demonstrated strong performance under domain shift, its behavior under class shift remains underexplored. To fill this gap, we evaluate WiSE-FT on 11 datasets in the base-to-new setting. Our results show that while WiSE-FT performs well on some datasets, it suffers significant performance degradation on others. Through our in-depth research, we identified a criterion capable of approximately assessing the utility of WiSE-FT on a given dataset. Specifically, we use the fine-tuned model to extract text features for the base and new classes, denoted as $\boldsymbol{w}_b$ and $\boldsymbol{w}_n$, respectively. Subsequently, we calculate the cosine distances both among the new class features and between the base and new class features:

$$\mathrm{D}_{n\_n} = avg(\mathbf{1} - \boldsymbol{w}_n \cdot \boldsymbol{w}_n^T - I), \mathrm{D}_{n\_b} = avg(\mathbf{1} - \boldsymbol{w}_b \cdot \boldsymbol{w}_n^T). \quad (15)$$

Finally, we define the difference between these two cosine distances as the adaptive criterion, denoted as $\Delta\mathrm{D} = \mathrm{D}_{n\_n} - \mathrm{D}_{n\_b}$. WiSE-FT is considered effective only when $\Delta\mathrm{D} \leq 0$.

As illustrated in Figure 3, to enable the use of WiSE-FT in prompt tuning, we follow the SkipTuning (Wu et al. 2025) strategy by unfreezing the last six layers of the visual encoder while freezing all other parameters, including

OOD Score	Train	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	Eurosat	UCF101	Average
MCM	✗	77.55	93.96	82.14	71.44	92.60	86.05	33.21	75.36	75.40	78.78	81.00	77.04
NegLabel	✗	95.96	88.72	92.97	95.08	89.33	91.96	76.39	77.47	68.53	85.39	84.81	86.06
GMoP	✓	97.30	99.40	98.07	96.00	95.85	97.15	68.34	90.10	80.16	84.34	93.38	90.92
ERF	✗	**98.70**	**99.77**	**99.25**	**98.38**	**97.91**	**98.43**	**89.82**	**94.10**	**83.41**	**90.94**	**97.11**	**95.26**

Table 1: OOD detection performance comparison with different methods. Our results are presented in bold.

the learnable ones introduced by the original PT method. The optimization is then performed using cross-entropy loss. At test time, samples identified by the OOD detector as belonging to base classes are directly classified using the fine-tuned model. For samples classified as new classes, we extract image features using the fine-tuned model if $\Delta \mathrm{D} > 0$. If $\Delta \mathrm{D} \leq 0$, we first merge the fine-tuned weights with the zero-shot weights and then use the merged model for feature extraction. Additionally, we use the class descriptions generated by GPT-3 (Brown et al. 2020), provided by CuPL (Pratt et al. 2023), as class templates to generate text features (denoted as $\boldsymbol{f}_l$), further enhancing the generalization ability of the classifier. Finally, we combine the prediction probabilities from both classifiers to obtain the final prediction:

$$\boldsymbol{p}\left(\boldsymbol{y}|\boldsymbol{x}\right) = \alpha \cdot \boldsymbol{p}\left(\boldsymbol{y}|\boldsymbol{f}_t\right) + (1-\alpha) \cdot \boldsymbol{p}\left(\boldsymbol{y}|\boldsymbol{f}_l\right), \tag{16}$$

where α controls the balance between the logits. We surprisingly find a synergistic effect between weight merge and logits merge, where the dual-merging significantly outperforms either method used alone (demonstrated by Figure 5).

Experiments

Datasets. We conduct experiments using 11 datasets from diverse sources. These datasets cover multiple recognition tasks including ImageNet (Deng et al. 2009) and Caltech101 (Fei-Fei, Fergus, and Perona 2004) for generic object recognition, OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Van Gool 2014) and FGVCAircraft (Maji et al. 2013) for fine-grained image recognition, EuroSAT (Helber et al. 2019) for satellite image classification, UCF101 (Soomro 2012) for action classification, DTD (Cimpoi et al. 2014) for texture classification, and SUN397 (Xiao et al. 2010) for scene recognition.

Compared Methods. We compare our method with two widely used training-free OOD detection approaches, MCM (Ming et al. 2022) and NegLabel (Jiang et al. 2024). Additionally, we include a comparison with GMoP (Hua et al. 2025), a recently proposed trainable method designed for the OPT task introduced by OpenworldAUC. The traditional PT methods we compared include PromptSRC (Khattak et al. 2023), TCP (Yao, Zhang, and Xu 2024), 2SFS (Farina et al. 2025), SkipTuning (Wu et al. 2025) and GMoP.

Evaluation Protocol. For OOD detection, we follow OpenworldAUC and report AUROC to measure the base-to-new class discriminability of different methods. AUROC is defined as the area under the curve of the true positive rate versus the false positive rate for the new super-class. For OPT, we report Opt-a, Opt-g, and Opt-h to assess the open-world recognition performance. For PT, we follow CoCoOp (Zhou et al. 2022a) and report Base, New, and H to evaluate the base-to-new generalization ability.

Implementation details. Following CoCoOp, we adopt a 16-shot prompt tuning setup using the pre-trained ViT-B/16 model from CLIP. To ensure fairness, all methods are reimplemented based on the open-source GitHub repository of SkipTuning. For each method, including GMoP, TCP, and SkipTuning, the optimizer, learning rate, batch size, and training epochs are kept consistent with their original implementations. For the learnable parameters in the last six layers of the visual encoder, we follow the training epochs and learning rate settings specified in SkipTuning. Both the weight merging coefficient and the logits merging coefficient are simply set to 0.5. Results are averaged over three runs with random seeds 1, 2, and 3.

Results

OOD Detection. For fair comparison, all OOD detection methods are applied to the trained SkipTuning model to compute AUROC scores, as reported in Table 1. Our method achieves the best performance across all 11 datasets, outperforming the non-trainable NegLabel by 9.20% and the trainable GMoP by 4.34%. These results demonstrate the effectiveness of our approach in enhancing base-to-new class discriminability.

Open-world Recognition. We evaluate the effectiveness of various OOD detection methods under three representative approaches: GMoP, TCP, and SkipTuning. As shown in Table 2, the same model exhibits substantial performance variation across different OOD detection methods, underscoring the importance of selecting an appropriate approach for the OPT task. Compared to classical OOD detection methods such as MCM and NegLable, as well as the training-based approach GMoP, our training-free method ERF consistently outperforms them on nearly all 11 datasets across three representative models, demonstrating its effectiveness and robustness. Moreover, integrating our proposed method GDM, which enhances model classification capability, improves the overall performance Opt-h of GMoP, TCP, and SkipTuning by 2.89%, 2.08%, and 1.41%, respectively. This demonstrates that even with a fixed OOD detection method, strengthening the model's classification ability can significantly boost performance.

Base-to-New Generalization. To assess the effectiveness of our plug-and-play method GDM in enhancing model adaptability and generalization, we incorporated it into the three models. The experimental results are presented in Table 3. Notably, our GDM significantly enhances the classification performance of all three models, with improvements ranging from 0.77% to 2.22% on base classes and 1.46% to 2.97% on novel classes. Furthermore, compared to conventional

Method	OOD Score	Average			ImageNet			Caltech101			OxfordPets		
		Opt-a	Opt-g	Opt-h	Opt-a	Opt-g	Opt-h	Opt-a	Opt-g	Opt-h	Opt-a	Opt-g	Opt-h
GMoP	MCM	68.07	57.54	53.05	62.73	51.52	42.90	92.03	87.55	86.54	82.71	82.51	80.44
	NegLabel	69.69	62.64	54.57	72.68	64.77	49.79	89.23	85.36	83.87	87.11	88.34	84.68
	GMoP	76.73	69.52	60.38	74.86	66.85	51.48	97.55	93.56	91.85	93.22	95.10	90.65
	ERF	**77.14**	**70.12**	**60.45**	**75.48**	**67.61**	**51.84**	**97.45**	**93.53**	**91.76**	**93.74**	**95.73**	**91.14**
	ERF+GDM	**79.17**	**72.09**	**63.34**	**76.53**	**70.50**	**54.82**	**97.61**	**94.55**	**92.91**	**93.80**	**95.95**	**91.41**
TCP	MCM	70.93	59.84	55.56	65.20	54.49	45.42	93.01	88.91	87.93	83.15	83.58	81.07
	NegLabel	72.47	64.63	56.99	74.07	66.63	51.74	89.51	86.05	84.62	87.96	89.96	85.67
	GMoP	78.11	69.98	61.84	75.63	68.12	53.01	97.78	94.04	92.53	92.62	95.31	90.26
	ERF	**80.13**	**71.96**	**63.07**	**76.31**	**68.96**	**53.43**	**98.06**	**94.37**	**92.79**	**93.84**	**96.71**	**91.42**
	ERF+GDM	**80.99**	**73.85**	**65.15**	**76.32**	**71.14**	**55.13**	**98.31**	**95.21**	**93.85**	**94.32**	**96.81**	**92.00**
SkipTuning	MCM	71.30	61.44	57.43	65.38	54.49	45.90	93.29	89.33	88.67	80.42	80.50	78.80
	NegLabel	74.11	67.69	59.87	75.28	67.50	52.95	87.52	84.52	83.38	89.10	91.15	87.36
	GMoP	79.21	72.13	64.12	76.55	68.76	54.04	98.14	94.59	93.38	93.88	96.02	91.91
	ERF	**81.45**	**74.71**	**65.75**	**77.26**	**69.57**	**54.44**	**98.45**	**94.89**	**93.64**	**94.88**	**97.18**	**92.90**
	ERF+GDM	**82.33**	**75.87**	**67.16**	**77.35**	**71.07**	**55.74**	**98.19**	**95.27**	**93.75**	**94.94**	**97.50**	**93.29**

Method	OOD Score	StanfordCars			Flowers102			Food101			FGVCAircraft		
		Opt-a	Opt-g	Opt-h	Opt-a	Opt-g	Opt-h	Opt-a	Opt-g	Opt-h	Opt-a	Opt-g	Opt-h
GMoP	MCM	55.34	49.04	41.69	88.75	69.20	68.14	77.04	75.12	70.23	19.50	9.16	6.18
	NegLabel	67.54	66.67	51.13	88.62	69.85	68.08	78.35	79.22	71.59	24.30	23.42	9.86
	GMoP	73.72	72.41	55.58	94.61	75.92	73.83	86.67	88.87	79.22	29.02	25.42	10.88
	ERF	**74.30**	**73.14**	**55.82**	**94.26**	**75.31**	**73.23**	**86.43**	**88.90**	**79.06**	**32.20**	**30.85**	**12.17**
	ERF+GDM	**79.86**	**74.43**	**61.08**	**95.31**	**75.08**	**73.83**	**86.71**	**89.60**	**79.96**	**39.22**	**34.95**	**16.79**
TCP	MCM	62.33	52.66	46.05	93.00	71.21	70.47	81.42	78.41	74.24	23.36	10.91	7.38
	NegLabel	76.79	70.36	56.78	91.77	70.75	69.39	83.65	83.07	76.43	31.34	25.75	11.36
	GMoP	77.68	71.31	57.68	95.30	74.19	72.68	88.67	88.88	81.06	30.57	24.44	11.04
	ERF	**79.52**	**72.99**	**58.75**	**96.69**	**75.05**	**73.48**	**89.36**	**89.85**	**81.71**	**36.49**	**30.74**	**12.97**
	ERF+GDM	**81.21**	**75.19**	**61.82**	**97.27**	**75.98**	**74.86**	**89.38**	**90.54**	**82.36**	**39.72**	**34.63**	**15.88**
SkipTuning	MCM	64.15	52.12	46.79	92.02	69.86	69.41	81.42	79.23	74.97	22.87	11.17	7.84
	NegLabel	78.82	69.65	57.73	88.04	67.65	66.68	84.77	84.88	78.21	33.93	30.59	13.71
	GMoP	79.99	70.51	58.65	94.46	72.82	71.74	88.75	89.80	81.97	33.47	26.90	12.92
	ERF	**81.49**	**72.00**	**59.62**	**96.34**	**74.54**	**73.32**	**89.55**	**90.97**	**82.71**	**39.40**	**34.15**	**15.17**
	ERF+GDM	**82.56**	**74.45**	**62.33**	**96.86**	**76.89**	**75.76**	**89.32**	**91.08**	**82.70**	**44.26**	**37.61**	**18.68**

Method	OOD Score	SUN397			DTD			Eurosat			UCF101		
		Opt-a	Opt-g	Opt-h	Opt-a	Opt-g	Opt-h	Opt-a	Opt-g	Opt-h	Opt-a	Opt-g	Opt-h
GMoP	MCM	62.78	53.92	46.92	62.74	42.92	37.22	74.35	49.73	47.44	70.80	62.27	55.85
	NegLabel	61.68	56.47	46.43	57.23	40.53	33.55	68.11	46.32	43.33	71.75	68.13	58.00
	GMoP	73.70	68.71	56.12	66.37	49.18	40.72	74.61	54.35	50.61	79.74	74.40	63.20
	ERF	**73.60**	**68.93**	**56.11**	**65.09**	**48.15**	**39.84**	**76.90**	**54.93**	**51.18**	**79.11**	**74.25**	**62.84**
	ERF+GDM	**74.50**	**71.57**	**59.00**	**67.76**	**48.56**	**41.86**	**77.87**	**62.07**	**58.80**	**81.74**	**75.71**	**66.29**
TCP	MCM	67.78	59.42	52.41	70.08	42.65	38.30	64.52	48.28	45.82	76.40	67.76	62.02
	NegLabel	65.61	60.75	51.18	59.96	38.35	33.03	62.09	49.02	44.69	74.47	70.30	62.01
	GMoP	75.79	70.81	59.53	69.09	45.71	39.41	73.31	60.56	55.48	82.83	76.42	67.52
	ERF	**78.05**	**73.43**	**61.38**	**71.12**	**47.61**	**41.07**	**76.93**	**63.00**	**57.49**	**85.06**	**78.92**	**69.32**
	ERF+GDM	**78.37**	**74.83**	**62.83**	**72.02**	**49.87**	**43.71**	**78.99**	**68.41**	**64.18**	**84.94**	**79.76**	**69.97**
SkipTuning	MCM	66.54	59.15	52.27	67.83	50.26	45.17	75.86	62.99	60.68	74.51	66.72	61.27
	NegLabel	65.17	61.27	51.54	58.85	46.24	39.66	79.23	69.59	64.56	74.54	71.51	62.79
	GMoP	75.67	72.07	60.41	69.44	55.32	47.71	78.28	69.36	64.34	82.64	77.30	68.28
	ERF	**78.17**	**75.36**	**62.50**	**71.14**	**57.88**	**49.43**	**84.13**	**74.77**	**69.12**	**85.12**	**80.46**	**70.43**
	ERF+GDM	**78.61**	**76.02**	**63.78**	**72.71**	**57.76**	**50.14**	**85.91**	**75.89**	**71.51**	**84.97**	**81.02**	**71.05**

Table 2: Performance comparison on eleven benchmark for open-world recognition. Our results are presented in bold.

prompt tuning methods such as DsRA and 2SFS, SkipTuning+GDM achieves state-of-the-art performance. This superior performance highlights DM's plug-and-play functionality, enabling seamless integration with existing methods.

Method	Avg. over 11 datasets		
	Base	New	H
PromptSRC	84.26	76.10	79.97
TCP	84.01	74.90	79.19
2SFS	85.55	75.48	80.20
SkipTuning	84.96	77.19	80.89
GMoP	82.21	74.33	78.07
TCP+GDM	**85.09**	**77.16**	**80.93**
SkipTuning+GDM	**85.73**	**78.65**	**82.04**
GMoP+GDM	**84.43**	**76.09**	**80.04**

Table 3: Base-to-new generalization experiments on eleven datasets. Our results are presented in bold.

	EuroSAT	UCF
D_{n_n} vs. D_{n_b}	$0.1055 > 0.0950$	$0.2850 > 0.2776$
w/o vs. w	$0.8065 > 0.7592$	$0.8222 > 0.8192$
	FGVC	Cars
D_{n_n} vs. D_{n_b}	$0.2298 < 0.2720$	$0.3529 < 0.3646$
w/o vs. w	$0.3679 < 0.3915$	$0.7269 < 0.7519$

Table 4: Explore the relationship between the distance among text features and the effectiveness of WiSE-FT. "w/o" and "w" denote the model's prediction accuracy on new classes without and with WiSE-FT, respectively.

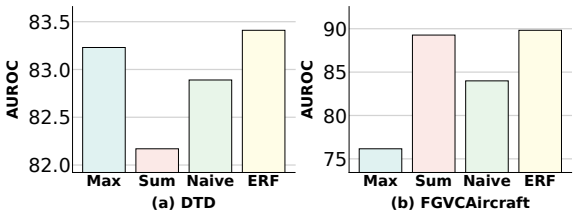

Figure 4: Ablation study of our ERF with SkipTuning as the baseline. "Naive" denotes the use of the default coefficient for mixup without rank normalization.

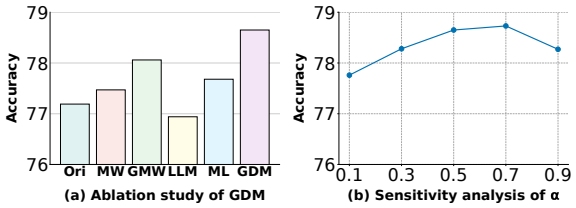

Figure 5: Ablation study and sensitivity analysis with SkipTuning as the baseline on eleven datasets. "Ori" denotes the original predictions from the SkipTuning method, "MW" denotes weight merging performed across all datasets, and "GMW" represents our gated merged weights. "LLM" refers to the results obtained by generating textual features using class descriptions produced by an LLM, which are then used for prediction. "ML" stands for merged logits.

Ablation Analysis

Metric for the effectiveness of WiSE-FT. Our experiments show that WiSE-FT is effective for new class prediction only when the distance between new class features is smaller than the distance between new and base class features, *i.e.*, when $\Delta D < 0$. Detailed results are presented in Table 4. We argue that if the distance between features of new classes exceeds that between new and base class features, it indicates that new classes are closer to base classes than to each other in feature space. This may cause feature space confusion, leading to interference from knowledge learned on base classes and ultimately reducing the accuracy on new classes.

Validity of our proposed ERF. We evaluate the effectiveness of our proposed OOD detection method, ERF, on the DTD and FGVCAircraft datasets using SkipTuning as the base model. As shown in Figure 4 (a) and (b), Max ($\boldsymbol{p}_m$) outperforms Sum ($\boldsymbol{p}_s$) on DTD, whereas $\boldsymbol{p}_s$ performs better on FGVCAircraft. A naive combination of the two yields suboptimal results. In contrast, ERF consistently achieves superior performance on both datasets, highlighting the robustness of our method.

Analysis of various merging combinations. Figure 5 (a) shows the ablation study of our GDM. The gated weight merging method outperforms the ungated version by 0.59%. This not only highlights the necessity of incorporating a gating mechanism into WiSE-FT, but also supports the validity of our proposed metric for evaluating the effectiveness of WiSE-FT. Compared to weight merging alone, our method improves performance by 0.58%, and by 0.97% over logits merging alone. These results indicate that weight and logits merging are complementary and can be effectively combined to enhance overall performance.

Impact of the balance coefficient α. The parameter α balances the contribution of fine-tuned and zero-shot logits. As shown in Figure 5 (b), performance peaks at $\alpha = 0.7$, with $\alpha = 0.5$ yielding comparable results.

Conclusion

This work systematically investigates open-world prompt tuning from three perspectives: evaluation metric, OOD detection, and optimization strategy. We introduce two novel metrics, Opt-a and Opt-g, to assess model adaptability and generalization, thereby completing the OPT evaluation framework. Additionally, we propose a training-free OOD detection method that dynamically adjusts the maximum and sum of base class probabilities using entropy weighting. We further investigate the performance of WiSE-FT under class distribution shift and propose an adaptive criterion to evaluate its effectiveness. Building on this, we introduce a plug-and-play dual-merging strategy that selectively merges the fine-tuned and zero-shot model at the weight level and combines fine-tuned and LLM-boosted logits at the output level. This strategy significantly enhances classification performance on both base and new classes. Extensive experiments on three representative PT baselines across eleven datasets demonstrate the effectiveness of our approach.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62176246. This work is also supported by Anhui Province Key Research and Development Plan (202304a05020045) and Anhui Province Natural Science Foundation (2208085UD17). This work is also supported by National Natural Science Foundation of China under Grant 62406098 and 62376256, and The Joint Fund for Medical Artificial Intelligence under Grant MAI2022Q011. This work is also supported by the Graduate School Special Funding Program of the University of Science and Technology of China.

References

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, 446–461. Springer.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dhamija, A. R.; Günther, M.; and Boult, T. 2018. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31.

Farina, M.; Mancini, M.; Iacca, G.; and Ricci, E. 2025. Rethinking Few-Shot Adaptation of Vision-Language Models in Two Stages. *arXiv preprint arXiv:2503.11609*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. 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*, 178–178. IEEE.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. 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.

Hua, C.; Xu, Q.; Yang, Z.; Wang, Z.; Bao, S.; and Huang, Q. 2025. OpenworldAUC: Towards Unified Evaluation and Optimization for Open-world Prompt Tuning. *arXiv preprint arXiv:2505.05180*.

Jiang, X.; Liu, F.; Fang, Z.; Chen, H.; Liu, T.; Zheng, F.; and Han, B. 2024. Negative label guided ood detection with pretrained vision-language models. *arXiv preprint arXiv:2403.20078*.

Khattak, M. U.; Wasim, S. T.; Naseer, M.; Khan, S.; Yang, M.-H.; and Khan, F. S. 2023. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15190–15200.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Ming, Y.; Cai, Z.; Gu, J.; Sun, Y.; Li, W.; and Li, Y. 2022. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35: 35087–35102.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 722–729. IEEE.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15691–15701.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15638–15650.

Soomro, K. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Wang, Z.; Xu, Q.; Yang, Z.; He, Y.; Cao, X.; and Huang, Q. 2022. Openauc: Towards auc-oriented open-set recognition. *Advances in Neural Information Processing Systems*, 35: 25033–25045.

Wortsman, M.; Ilharco, G.; Kim, J. W.; Li, M.; Kornblith, S.; Roelofs, R.; Lopes, R. G.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7959–7971.

Wu, S.; Zhang, J.; Zeng, P.; Gao, L.; Song, J.; and Shen, H. T. 2025. Skip tuning: Pre-trained vision-language models are effective and efficient adapters themselves. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14723–14732.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Xiao, K.; Wang, Z.; and Li, J. 2024. Semantic-guided robustness tuning for few-shot transfer across extreme domain shift. In *European Conference on Computer Vision*, 303–320. Springer.

Yang, Z.; Xu, Q.; Bao, S.; Cao, X.; and Huang, Q. 2021. Learning with multiclass AUC: Theory and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7747–7763.

Yao, H.; Zhang, R.; and Xu, C. 2024. TCP: Textual-based Class-aware Prompt tuning for Visual-Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23438–23448.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, Z.; Yang, M.; Shi, J.-X.; Guo, L.-Z.; and Li, Y.-F. 2024. DeCoOp: robust prompt tuning with out-of-distribution detection. *arXiv preprint arXiv:2406.00345*.