# Bi-CoG: Bi-Consistency-Guided Self-Training for Vision-Language Models

**Rui Zhu**[1] , **Song-Lin Lv**[1,2] , **Zi-Kang Wang**[1,2] and **Lan-Zhe Guo**[1,2*]
[1]School of Intelligence Science and Technology, Nanjing University, China
[2]National Key Laboratory for Novel Software Technology, Nanjing University, China
zhurui@smail.nju.edu.cn, {lvsl,wangzk,guolz}@lamda.nju.edu.cn

## Abstract

Exploiting unlabeled data through semi-supervised learning (SSL) or leveraging pre-trained models via fine-tuning are two prevailing paradigms for addressing label-scarce scenarios. Recently, growing attention has been given to combining fine-tuning of pre-trained vision-language models (VLMs) with SSL, forming the emerging paradigm of semi-supervised fine-tuning. However, existing methods often suffer from model bias and hyperparameter sensitivity, due to reliance on prediction consistency or pre-defined thresholds. To address these limitations, we propose a simple yet effective plug-and-play method named **Bi-Co**nsistency-**G**uided Self-Training (Bi-CoG), which assigns accurate and low-bias pseudo-labels, by simultaneously exploiting inter- and intra-model consistency, along with an error-aware dynamic filtering strategy. Both theoretical analysis and extensive experiments over 14 datasets demonstrate the effectiveness of Bi-CoG, which consistently and significantly improves the performance of existing methods. Codes are available at Bi-CoG.

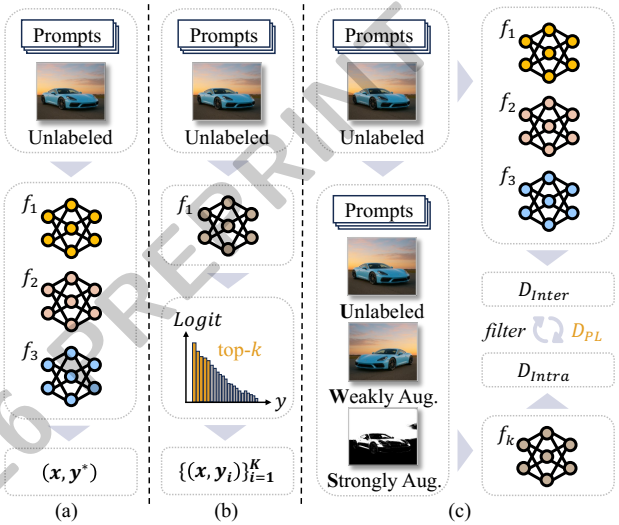



Figure 1: Comparison of existing pseudo-labeling paradigms and Bi-CoG. (a) Majority voting from multiple VLMs; (b) Top-$k$ selection based on predicted logits; (c) Bi-CoG dynamically generates pseudo-labels based on bi-consistency and model error.

## 1 Introduction

Addressing downstream tasks with limited labeled data remains a critical challenge in real-world applications [Yang *et al.*, 2023; Shao *et al.*, 2026; Yang *et al.*, 2025]. Two prominent paradigms have emerged to tackle this issue: (1) adapting pre-trained models, such as vision-language models (VLMs) [Radford *et al.*, 2021; Jia *et al.*, 2021], for downstream tasks; and (2) applying semi-supervised learning (SSL) methods that leverage large amounts of inexpensive unlabeled data. Recent work [Lv *et al.*, 2025b] has further suggested combining them, where the strong zero-shot capabilities of pre-trained VLMs are used to generate reliable pseudo-labels that guide semi-supervised fine-tuning, thereby fully exploiting the strengths of both paradigms.

Recent efforts have explored integrating these two paradigms by leveraging pre-trained models to extract domain knowledge from unlabeled data for downstream task adaptation. A common approach is model ensembling (Figure 1.a), such as V-PET [Gupte *et al.*, 2024], which aggregates multi-view one-hot predictions from multiple vision models. As shown in Figure 1.b, another line of work selects top-$k$ predictions using class-wise or confidence-based thresholds. For instance, GRIP [Menghini *et al.*, 2023] and UPL [Huang *et al.*, 2022] enforce class balance by selecting equal samples per class; CPL [Zhang *et al.*, 2024] refines labels via instance-level confidence matrices.

However, these methods struggle to balance the trade-off between pseudo-label accuracy and bias. Ensemble-based approaches can produce highly accurate pseudo-labels, but they often inadvertently amplify biases inherited from pre-training, leading to uneven performance across subgroups—a phenomenon commonly referred to as the *"Matthew effect"* [Zhu *et al.*, 2022; Chen *et al.*, 2022; Gan and Wei, 2024; Guo *et al.*, 2020]. In contrast, soft-labeling techniques and class-wise pseudo-label allocation strategies can alleviate overconfidence to some extent, but they tend to introduce additional label noise and are highly sensitive to manually tuned

*Corresponding author.

hyperparameters. To the best of our knowledge, adaptively generating reliable and unbiased pseudo-labels in response to model performance remains a critical yet underexplored challenge in semi-supervised fine-tuning.

To address this challenge, we propose a novel method named **Bi-Co**nsistency-**G**uided Self-Training (Bi-CoG). As illustrated in Figure 1.c, Bi-CoG leverages multiple VLMs as base learners to perform dynamic self-training guided by both inter- and intra-model consistency as well as model error estimation. Specifically, we conduct a dynamic self-training process that comprises three-stage pseudo-label selection. In the first stage, we apply majority voting to generate ensembled pseudo-labels, ensuring reliable supervision. Then, we assess each model's bias on a sample using weak and strong augmentations: predictions consistent under no and weak augmentation but changing under strong augmentation indicate reliability and less prone to pre-training bias. Finally, based on theoretical insights, we dynamically adjust the upper bound on the number of pseudo-labels used for training, according to models' relative accuracy on labeled data. This further filters pseudo-labels and enables adaptive self-training throughout the training process.

As a plug-and-play approach, Bi-CoG can be seamlessly integrated into existing VLM fine-tuning pipelines. To evaluate its effectiveness, we conduct experiments under both *open-world generalization* [Zhou *et al.*, 2024] and *standard semi-supervised learning*[Yang *et al.*, 2023] settings, combining Bi-CoG with various prompt-tuning methods. Extensive results demonstrate the effectiveness of our bi-consistency-guided dynamic self-training strategy, which consistently improves performance across 14 datasets without relying on larger models or external knowledge, outperforming state-of-the-art (SOTA) methods.

Our main contributions are summarized as follows:

- We propose Bi-CoG, a plug-and-play method that jointly leverages the pre-trained knowledge of VLMs and external information from unlabeled data.
- We design a deeply coupled self-training strategy that integrates inter-model consistency, intra-model consistency and model error estimation, enabling reliable, low-bias pseudo-label generation and adaptive training.
- Theoretical analysis and extensive experiments demonstrate that Bi-CoG consistently improves performance, achieving up to a 5.69% gain across 14 datasets and surpassing SOTA methods.

## 2 Related Work

**Parameter-efficient Fine-tuning on VLMs** While pre-trained VLMs demonstrate strong zero-shot generalization in domain-specific downstream tasks, their large parameter sizes often hinder efficient adaptation to specific applications [Zhou *et al.*, 2022b]. To address this, recent works have proposed parameter-efficient fine-tuning approaches that introduce a small number of trainable parameters while keeping the majority of the model frozen [Li *et al.*, 2024]. Tip-Adapter [Zhang *et al.*, 2022] and CLIP-Adapter [Gao *et al.*, 2024] append lightweight bottleneck layers after the text or image encoders to learn residual features for better representation. CoOp [Zhou *et al.*, 2022b] introduces prompt tuning to VLMs for the first time by replacing manually crafted hard prompts (e.g., "a photo of a classname") with learnable soft text prompts optimized via backpropagation. Building on this idea, methods such as MaPLe [Khattak *et al.*, 2023a], PromptSRC [Khattak *et al.*, 2023b], and BMIP [Lv *et al.*, 2025a] extend prompt tuning to both visual and textual encoders, enabling joint vision-language adaptation. While these methods have advanced the application of VLMs to downstream tasks, the transformation of prompts into continuous vectors makes them prone to overfitting and leads to sub-optimal performance in open-world scenarios [Zhou *et al.*, 2022a; Khattak *et al.*, 2023b]. Beyond prompt tuning, recent work on VLM selection and reuse [Tan *et al.*, 2025] further explores how to choose and adapt pre-trained models for downstream tasks.

**Pseudo-labeling in Semi-supervised Learning** Semi-supervised learning (SSL) addresses label scarcity by leveraging both labeled and unlabeled data. Pseudo-labeling, the most widely used SSL approach, assigns labels to unlabeled data based on model predictions and improves performance through self-training. The key to effective pseudo-labeling lies in the generation of high-quality pseudo-labels [Yang *et al.*, 2023]. Traditional SSL methods typically filter pseudo-labels using confidence thresholds or adopt data augmentation to generate diverse views, aiming to improve label quality and model robustness [Lee and others, 2013]. For VLMs, their strong generalization ability allows them to produce highly accurate pseudo-labels even in zero- or few-shot settings, making them strong candidates as pseudo-labelers. Recent works have explored more advanced strategies: GRIP [Menghini *et al.*, 2023] gradually increases the upper limit on the number of pseudo-labels to mitigate calibration errors [Oh *et al.*, 2024] and label imbalance [Wang *et al.*, 2022; Guo and Li, 2022]; and CPL [Zhang *et al.*, 2024] progressively refines pseudo-labels based on a confidence score matrix to avoid overconfidence. Despite their effectiveness, these methods struggle with balancing pseudo-label noise and inherited pre-training biases [Oh *et al.*, 2024; Wang *et al.*, 2022; Li *et al.*, 2021], while failing to dynamically adapt to model performance in open environments [Guo *et al.*, 2025], ultimately limiting their potential.

**Problem Definition** Based on the aforementioned challenges in pseudo-labeling—specifically the issues of label noise and static adaptation—we formally define our task. In downstream scenarios with scarce labeled data, we typically have access to a small labeled dataset denoted as $D_L = \{(x_i, y_i)\}_{i=1}^{M}$, along with a large amount of unlabeled data $D_{UL} = \{x_i\}_{i=1}^{N}$, where $M \ll N$. Given a collection of $K$ pre-trained VLMs $\{f_k\}_{k=1}^{K}$ (with $K \geq 3$), the objective is to effectively leverage both $D_L$ and $D_{UL}$ to fine-tune the models and maximize performance on the target dataset $D_{test}$.

## 3 Methodology

To address the prevalent challenges of balancing pseudo-label accuracy with model bias and reducing the strong reliance on empirical parameter tuning in prior methods, we propose

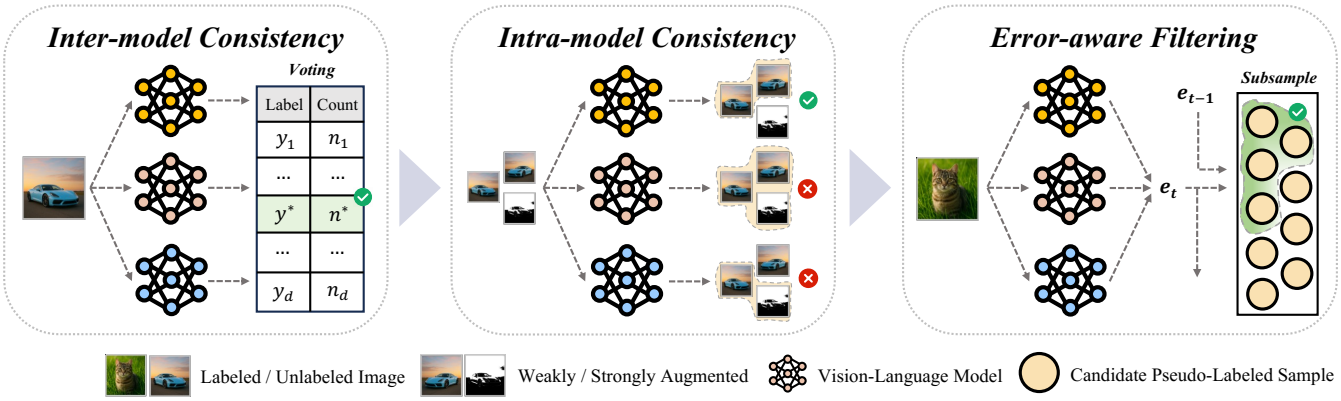



Figure 2: The overall framework of Bi-CoG's pseudo-label selection. During the self-training phase, Bi-CoG employs inter-model consistency, intra-model consistency, and error-aware filtering to generate accurate and unbiased pseudo-labels.

Bi-Consistency-Guided Self-Training (Bi-CoG), a simple yet effective plug-and-play framework. As illustrated in Figure 2, the Bi-CoG framework comprises the following three major components for pseudo-label selection:

- A model ensembling module that aggregates predictions through majority voting based on inter-model consistency, generating high-quality candidate pseudo-labels to ensure reliable supervision.
- A multi-view filtering mechanism that leverages intra-model consistency by applying diverse transformations to unlabeled data and discarding biased samples with inconsistent predictions, thereby mitigating the influence of pre-training bias.
- An error-aware bounding strategy that dynamically estimates model performance and sets an upper limit on the number of pseudo-labels used, promoting stable and progressive self-training.

The procedure of Bi-CoG is summarized in Algorithm 1. We elaborate on each component in the subsequent sections.

**Algorithm 1** Bi-CoG's overall learning algorithm.

**Input**: Pre-trained VLMs $\{f_k\}_{k=1}^K$, labeled data $D_L$, unlabeled data $D_{UL}$, scaling factor $\alpha$
**Output**: Fine-tuned VLMs

**for** $j = 1$ **to** $K$ **do**
  $e'_j \leftarrow 0.5$, update$_j \leftarrow$ False
**end for**
$t \leftarrow 0$, improve $\leftarrow$ True
**while** improve **do**
  $t \leftarrow t + 1$
  **for** $j = 1$ **to** $K$ **do**
    update$_j \leftarrow$ False
    $e_j \leftarrow \text{MeasureError}(D_L, \{f_k\}_{k\neq j}^K)$
    **if** $e_j < e'_j$ **then**
      $D_{Inter}^j \leftarrow \text{InterConsistency}(D_{UL}, \{f_k\}_{k\neq j}^K)$
      $D_{Intra}^j \leftarrow \text{IntraConsistency}(D_{UL}, \{f_k\}_{k\neq j}^K)$
      update$_j \leftarrow \text{ErrorAware}\left(e_j,\ e'_j,\ t\right)$
      $D_{PL}^j \leftarrow \text{Subsample}(D_{Inter}^j \cap D_{Intra}^j, (\frac{e'_j}{e_j})^{\alpha t} L'_j)$
    **end if**
    **if** update$_j$ **then**
      $e'_j \leftarrow e_j,\ f_j \leftarrow \text{FineTune}(f_j, D_L \cup D_{PL}^j)$
    **end if**
  **end for**
  **if** update $==$ [False] $* K$ **then**
    improve $\leftarrow$ False
  **end if**
**end while**
**return** $\{f_k\}_{k=1}^K$

## 3.1 Inter-model Consistency

During self-training with pseudo-labels, label fidelity is critical to effective and stable optimization [Angluin and Laird, 1988]. Excessive noise can mislead learning, causing convergence to suboptimal representations. To mitigate this issue, we first generate pseudo-labels based on inter-model consistency by integrating predictions of multiple VLMs, with the primary goal of ensuring the accuracy of pseudo-labels.

Concretely, for each model $f_j$ $(1 \leq j \leq K)$, we construct a candidate pseudo-labeled dataset $D_{Inter}^j$ by selecting unlabeled instances on which the majority of the other $K-1$ models reach a consensus. For each unlabeled sample $x_i \in D_{UL}$, we first compute the class-wise vote count from the other models as follows:

$$v_i(y) = \sum_{1\leq k\leq K,\ k\neq j} \mathbb{I}\left[f_k(x_i) = y\right] \tag{1}$$

If the most voted label $y^* = \arg\max_{y\in\mathcal{Y}} v_i(y)$ satisfies $v_i(y^*) \geq \lceil \frac{K-1}{2} \rceil$, we assign $x_i$ a pseudo-label $\hat{y}_i = y^*$ and include the pair $(x_i, y^*)$ in $D_{Inter}^j$. Otherwise, the sample is discarded. Formally, this candidate set is defined as:

$$D_{Inter}^j = \left\{ (x_i, y^*) \,\middle|\, v_i(y^*) \geq \left\lceil \frac{K-1}{2} \right\rceil \right\} \tag{2}$$

where $\mathcal{Y}$ denotes the label space, and $\mathbb{I}[\cdot]$ is the indicator function. This strategy ensembles multiple models to generate reliable pseudo-labels, reducing noise from individual prediction errors. Moreover, the leave-one-out scheme prevents amplifying biases inherent to the target model, thereby resulting in more accurate and unbiased data for fine-tuning.

### 3.2 Intra-model Consistency

While inter-model consistency improves pseudo-label reliability by aggregating agreement across models, it may still preserve instances where VLMs share similar biases, often due to overlapping pre-training corpora or architectural homogeneity [Gan and Wei, 2024; Wang and Russakovsky, 2023]. To further mitigate biased supervision, we introduce intra-model consistency, which leverages the consistency of a single model's predictions under different input views to refine the pseudo-labeled sets obtained in the previous stage.

Specifically, for each unlabeled sample $x_i \in D_{UL}$, we apply both weak and strong data augmentations. Each model $f_k$ then produces predictions on the original input $x_i$, the weakly augmented $\tilde{x}_i^w$, and the strongly augmented $\tilde{x}_i^s$. A sample is retained only if the prediction remains consistent between the original and weakly augmented inputs, i.e., $f_k(x_i) = f_k(\tilde{x}_i^w)$, but differs under strong augmentation, i.e., $f_k(\tilde{x}_i^s) \neq f_k(x_i)$. This criterion identifies samples with low uncertainty and meaningful sensitivity, as evidenced by prediction consistency under weak perturbations and inconsistency under strong perturbations. The intra-model consistent pseudo-label set for $f_j$ is defined as the intersection of predictions from the other models:

$$D_{Intra}^j = \bigcap_{\substack{k=1 \\ k\neq j}}^{K} \left\{ (x_i, f_k(x_i)) \middle| \begin{array}{l} f_k(x_i) = f_k(\tilde{x}_i^w) \\ f_k(x_i) \neq f_k(\tilde{x}_i^s) \end{array} \right\} \quad (3)$$

Once the $D_{Intra}^j$ are computed, the candidate pseudo-label set is constructed by intersecting its inter-model consistent set $D_{Inter}^j$ with the intra-model consistent set:

$$D_{PL}^j = D_{Inter}^j \cap D_{Intra}^j \quad (4)$$

The intersection of pseudo-labels from inter-model consistency and intra-model consistency preserves accuracy while mitigating shared model biases. Consequently, the resulting pseudo-labeled datasets are both highly reliable and low in bias, providing a robust foundation for the subsequent dynamic self-training stage.

### 3.3 Error-aware Filtering

After bi-consistency-guided pseudo-labeling, we obtain a candidate sample set $D_{PL}$, which contains high-accuracy and low-biased pseudo-labels. However, during self-training, it is crucial to determine an appropriate upper bound on the number of pseudo-labeled samples based on model performance. Prior studies [Angluin and Laird, 1988; Chen *et al.*, 2022] have shown that even minor label noise, if not aligned with model competence, can significantly degrade performance, making further noise reduction in $D_{PL}$ essential. Existing methods [Menghini *et al.*, 2023; Huang *et al.*, 2022] typically rely on fixed thresholds to limit the number of pseudo-labels, which can be suboptimal as model performance evolves over time and across different datasets.

To address this limitation, we dynamically adjust the upper bound on pseudo-label usage based on estimated model error ratio, guided by a theoretical analysis. Specifically, following prior work [Angluin and Laird, 1988; Zhou and Li, 2005] we formalize the influence of noisy supervision on model $f_j$ during iterative self-training:

**Lemma 1.** *Let $L_t$ denote the training set $D_{train}$ for model $f_j$ at iteration $t$, and $e_t$ represent the relative error rate of the remaining $K-1$ models at the same iteration, i.e., the probability that the majority vote produces an incorrect label. Then, training $f_j$ on $D_{train}$ will result in performance improvement if the following condition is satisfied:*

$$0 < \frac{e_t}{e_{t-1}} < \frac{L_{t-1}}{L_t} < 1 \quad (5)$$

According to Lemma 1, the model will be improved throughout the training process as long as Equation 5 holds. In practice, we estimate the theoretical error rate ratio from the performance of the $K-1$ models on the labeled set. However, this estimation can be biased, as repeated training on labeled data tends to produce stable—and often low—error ratios, which are reflected in Equation 5 as values close to 1. In contrast, the actual error ratio on unlabeled data typically decreases during training, making the true ratio $\frac{e_t}{e_{t-1}}$ lower than the estimated $\frac{\hat{e}_t}{\hat{e}_{t-1}}$. Ignoring this discrepancy would introduce bias into the estimated pseudo-label budget, compromising the validity of the theoretical analysis. To address this issue, we introduce Theorem 1 to provide a correction:

**Theorem 1.** *As the training progresses, the true error rate ratio can be approximately equal to the $t$-power of the estimated error rate ratio, as shown below:*

$$\frac{e_t}{e_{t-1}} \approx \left[\frac{\hat{e}_t}{\hat{e}_{t-1}}\right]^{\alpha t} \quad (6)$$

*where $\alpha > 0$ is a scaling factor that reflects the error reduction rate. Then, the original condition in Equation 5 is approximately equivalent to the following constraints:*

$$\begin{aligned} L_t &\leq \left\lceil \left(\frac{\hat{e}_{t-1}}{\hat{e}_t}\right)^{\alpha t} \times L_{t-1} - 1 \right\rceil \\ L_{t-1} &> \frac{\hat{e}_t^{\alpha t}}{\hat{e}_{t-1}^{\alpha t} - \hat{e}_t^{\alpha t}} \end{aligned} \quad (7)$$

Based on the theoretical analysis, we establish the necessary condition under which the model can be effectively optimized by self-training, as described in Equation 7, and establish an upper bound on the pseudo-label budget for each iteration. Here, $L_j'$ denotes the size of the pseudo-labeled training set used by model $f_j$ in the previous iteration. Accordingly, the final set of pseudo-labeled samples is formalized as:

$$D_{PL^*}^j = \text{Subsample}\left(D_{PL}^j, \left(\frac{e_j'}{e_j}\right)^{\alpha t} L_j'\right) \quad (8)$$

This dynamic filtering strategy reduces the mismatch between model capacity and pseudo-label quantity. Consequently, the model can self-regulate the amount of pseudo-labeled data during training, leading to stable and consistent performance improvements.

| Method | Average | | ImageNet | | Caltech101 | | OxfordPets | | StanfordCars | |
|---|---|---|---|---|---|---|---|---|---|---|
| | HM | Acc. | HM | Acc. | HM | Acc. | HM | Acc. | HM | Acc. |
| CLIP | 73.00 | 66.41 | 70.18 | 66.70 | 95.68 | 93.30 | 94.11 | 89.10 | 68.85 | 65.60 |
| CoOp | 72.10 | 66.59 | 72.00$_{\pm 0.34}$ | 68.73$_{\pm 0.25}$ | 90.91$_{\pm 1.61}$ | 89.90$_{\pm 0.99}$ | 93.97$_{\pm 0.72}$ | 88.73$_{\pm 0.09}$ | 69.94$_{\pm 0.59}$ | 66.40$_{\pm 0.50}$ |
| + *Bi-CoG* | **77.79** | **71.03** | **73.06**$_{\pm 0.08}$ | **69.78**$_{\pm 0.04}$ | **96.67**$_{\pm 0.05}$ | **94.24**$_{\pm 0.25}$ | **96.28**$_{\pm 0.60}$ | **89.57**$_{\pm 2.02}$ | **73.38**$_{\pm 0.31}$ | **69.31**$_{\pm 0.76}$ |
| MaPLe | 78.78 | 72.18 | 73.58$_{\pm 0.09}$ | 70.30$_{\pm 0.14}$ | 96.58$_{\pm 0.21}$ | 94.87$_{\pm 0.09}$ | 96.60$_{\pm 0.36}$ | 92.33$_{\pm 0.31}$ | 73.53$_{\pm 0.81}$ | 69.97$_{\pm 0.87}$ |
| + *Bi-CoG* | **79.77** | **73.54** | **74.11**$_{\pm 0.07}$ | **70.82**$_{\pm 0.09}$ | **96.77**$_{\pm 0.17}$ | **95.09**$_{\pm 0.46}$ | **96.93**$_{\pm 0.06}$ | **92.56**$_{\pm 0.12}$ | **74.20**$_{\pm 0.31}$ | **70.48**$_{\pm 0.36}$ |
| PromptSRC | 79.99 | 74.11 | 74.07$_{\pm 0.06}$ | 70.77$_{\pm 0.05}$ | 96.11$_{\pm 0.08}$ | 94.63$_{\pm 0.12}$ | 96.26$_{\pm 0.29}$ | 92.13$_{\pm 0.59}$ | 76.68$_{\pm 0.16}$ | 73.03$_{\pm 0.19}$ |
| + *Bi-CoG* | **81.46** | **74.96** | **74.18**$_{\pm 0.03}$ | **70.94**$_{\pm 0.03}$ | **96.45**$_{\pm 0.16}$ | **94.74**$_{\pm 0.16}$ | **96.37**$_{\pm 0.13}$ | **92.62**$_{\pm 0.13}$ | **79.07**$_{\pm 0.15}$ | **73.77**$_{\pm 0.01}$ |

| Method | Flowers102 | | Food101 | | FGVCAircraft | | SUN397 | | DTD | |
|---|---|---|---|---|---|---|---|---|---|---|
| | HM | Acc. | HM | Acc. | HM | Acc. | HM | Acc. | HM | Acc. |
| CLIP | 74.44 | 70.70 | 90.39 | 85.60 | 31.21 | 24.80 | 72.37 | 62.60 | 56.62 | 44.00 |
| CoOp | 75.70$_{\pm 2.82}$ | 71.27$_{\pm 2.21}$ | 85.82$_{\pm 1.41}$ | 80.03$_{\pm 1.17}$ | 29.93$_{\pm 1.54}$ | 26.07$_{\pm 0.74}$ | 71.60$_{\pm 1.81}$ | 63.33$_{\pm 1.44}$ | 55.93$_{\pm 1.25}$ | 50.87$_{\pm 0.62}$ |
| + *Bi-CoG* | **83.57**$_{\pm 0.91}$ | **79.59**$_{\pm 1.22}$ | **90.54**$_{\pm 0.14}$ | **85.03**$_{\pm 0.65}$ | **35.61**$_{\pm 0.29}$ | **28.40**$_{\pm 0.21}$ | **76.89**$_{\pm 0.06}$ | **67.15**$_{\pm 0.18}$ | **59.90**$_{\pm 1.37}$ | **50.96**$_{\pm 1.35}$ |
| MaPLe | 82.98$_{\pm 0.19}$ | 78.47$_{\pm 0.60}$ | 90.83$_{\pm 0.35}$ | 86.47$_{\pm 0.40}$ | 35.30$_{\pm 0.88}$ | 27.30$_{\pm 0.96}$ | 79.42$_{\pm 0.30}$ | 70.97$_{\pm 0.24}$ | 67.40$_{\pm 3.09}$ | 55.30$_{\pm 1.40}$ |
| + *Bi-CoG* | **84.62**$_{\pm 0.25}$ | **79.12**$_{\pm 0.51}$ | **91.67**$_{\pm 0.03}$ | **87.46**$_{\pm 0.02}$ | **35.58**$_{\pm 0.71}$ | **27.63**$_{\pm 0.56}$ | **80.27**$_{\pm 0.19}$ | **71.82**$_{\pm 0.23}$ | **69.88**$_{\pm 1.07}$ | **56.36**$_{\pm 0.73}$ |
| PromptSRC | 86.13$_{\pm 0.39}$ | 80.77$_{\pm 0.50}$ | 90.91$_{\pm 0.12}$ | 86.47$_{\pm 0.12}$ | 34.99$_{\pm 6.68}$ | 29.93$_{\pm 2.50}$ | 80.68$_{\pm 0.23}$ | 72.13$_{\pm 0.26}$ | 70.72$_{\pm 1.23}$ | 58.87$_{\pm 1.28}$ |
| + *Bi-CoG* | **87.60**$_{\pm 0.18}$ | **83.19**$_{\pm 0.55}$ | **91.25**$_{\pm 0.01}$ | **86.65**$_{\pm 0.02}$ | **39.89**$_{\pm 0.34}$ | **31.98**$_{\pm 0.69}$ | **80.89**$_{\pm 0.04}$ | **72.43**$_{\pm 0.10}$ | **71.67**$_{\pm 0.42}$ | **59.48**$_{\pm 0.10}$ |

| Method | EuroSAT | | UCF101 | | CIFAR-10 | | CIFAR-100 | | ImageNet-100 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | HM | Acc. | HM | Acc. | HM | Acc. | HM | Acc. | HM | Acc. |
| CLIP | 60.21 | 48.30 | 74.42 | 67.50 | 88.43 | 79.00 | 55.17 | 46.00 | 88.33 | 86.50 |
| CoOp | 72.04$_{\pm 3.29}$ | 59.50$_{\pm 2.10}$ | 67.03$_{\pm 1.48}$ | 64.23$_{\pm 1.26}$ | 89.17$_{\pm 1.32}$ | 77.00$_{\pm 2.42}$ | 53.42$_{\pm 3.23}$ | 45.03$_{\pm 2.88}$ | 81.97$_{\pm 2.38}$ | 81.17$_{\pm 1.96}$ |
| + *Bi-CoG* | **85.84**$_{\pm 1.98}$ | **67.98**$_{\pm 0.67}$ | **73.11**$_{\pm 1.99}$ | **68.01**$_{\pm 0.92}$ | **92.07**$_{\pm 0.46}$ | **82.28**$_{\pm 0.99}$ | **61.71**$_{\pm 0.26}$ | **52.55**$_{\pm 0.32}$ | **90.39**$_{\pm 0.35}$ | **89.59**$_{\pm 0.30}$ |
| MaPLe | 78.35$_{\pm 1.46}$ | 60.90$_{\pm 8.71}$ | 81.57$_{\pm 0.39}$ | 74.70$_{\pm 0.57}$ | 92.02$_{\pm 0.76}$ | 83.83$_{\pm 1.21}$ | 64.73$_{\pm 0.30}$ | 56.10$_{\pm 0.37}$ | 90.00$_{\pm 0.17}$ | 89.07$_{\pm 0.33}$ |
| + *Bi-CoG* | **81.33**$_{\pm 2.89}$ | **70.52**$_{\pm 2.81}$ | **82.23**$_{\pm 0.58}$ | **75.75**$_{\pm 0.51}$ | **92.15**$_{\pm 0.09}$ | **84.41**$_{\pm 0.69}$ | **66.68**$_{\pm 0.52}$ | **57.97**$_{\pm 0.61}$ | **90.41**$_{\pm 0.29}$ | **89.52**$_{\pm 0.32}$ |
| PromptSRC | 81.87$_{\pm 2.52}$ | 70.03$_{\pm 0.86}$ | 82.30$_{\pm 0.87}$ | 76.67$_{\pm 0.37}$ | 91.46$_{\pm 0.30}$ | 83.80$_{\pm 0.42}$ | 67.23$_{\pm 0.42}$ | 58.93$_{\pm 0.39}$ | 90.49$_{\pm 0.22}$ | 89.43$_{\pm 0.26}$ |
| + *Bi-CoG* | **86.28**$_{\pm 0.87}$ | **71.01**$_{\pm 0.47}$ | **83.09**$_{\pm 0.24}$ | **77.30**$_{\pm 0.44}$ | **94.50**$_{\pm 0.15}$ | **84.96**$_{\pm 1.75}$ | **68.19**$_{\pm 0.38}$ | **60.20**$_{\pm 0.07}$ | **91.07**$_{\pm 0.02}$ | **90.15**$_{\pm 0.05}$ |

Table 1: Performance comparison on 14 datasets using ViT-B/16 architecture. The best performance is in bold.



After training, Bi-CoG adopts a majority voting strategy to generate predicted labels for evaluation during the testing phase, ensuring robust and reliable assessment results.

In summary, Bi-CoG offers the following key advantages:

1. **Plug-and-Play Adaptability**: Bi-CoG can be seamlessly integrated with existing pre-trained models, providing consistent performance improvements with minimal changes to the original method.
2. **Robust Pseudo-Labeling**: By jointly leveraging bi-consistency and error-aware mechanisms, Bi-CoG generates more reliable pseudo-labels, effectively reducing label noise and alleviating model bias.
3. **Adaptive Self-Training Control**: Empowered by the error-aware mechanism, Bi-CoG can automatically regulate the self-training process, eliminating the need for manually preset iteration counts and enhances adaptability across different scenarios.

# 4 Experiments

In this section, we design experiments to answer the following research questions:

1. **RQ1:** Does Bi-CoG enhance existing PEFT methods for VLMs in open-world generalization scenarios?
2. **RQ2:** Can Bi-CoG deliver competitive performance compared to conventional SSL approaches?
3. **RQ3:** How does Bi-CoG compare to other methods that jointly leverage pre-trained models and unlabeled data?
4. **RQ4:** What is the contribution of each component within the Bi-CoG framework?

## 4.1 Experimental Settings

**Evaluation Paradigm.** To answer RQ1 and RQ2, we evaluate the effectiveness of Bi-CoG on two widely adopted settings: *open-world generalization setting* [Zhou *et al.*, 2024;

Lv *et al.*, 2025a] and *semi-supervised setting* [Lee and others, 2013; Yang *et al.*, 2023]. In *open-world generalization setting*, the model is trained only on a subset of categories (referred to as base classes), while the remaining categories are treated as novel classes. We report the harmonic mean (HM) of the model's performance on base and novel classes, as well as the overall accuracy on the full test set, which includes both base and novel classes. *Semi-supervised setting* follows a common protocol where only a small portion of the data is labeled, while the majority remains unlabeled.

**Dataset and Baselines.** We evaluate the performance of our method on 14 recognition datasets, including 11 commonly used for assessing VLMs and 3 classical benchmarks for semi-supervised learning. We select CLIP [Radford *et al.*, 2021] and its three representative prompt tuning methods as baselines: CoOp [Zhou *et al.*, 2022b], MaPLe [Khattak *et al.*, 2023a], and PromptSRC [Khattak *et al.*, 2023b], among which PromptSRC represents the current SOTA method. For methods that integrate the two paradigms, we select GRIP [Menghini *et al.*, 2023] and CPL [Zhang *et al.*, 2024] as baselines, with CPL representing the current SOTA approach.

**Implementation Details.** We set the number of VLMs $K = 3$ throughout all experiments. The setting $K \geq 3$ ensures that an absolute majority can be achieved in most cases, facilitating reliable pseudo-label selection. While a larger $K$ further improves stability, it also increases GPU memory consumption. We adopt $K = 3$ to balance effectiveness and efficiency. For the scaling factor $\alpha$, it provides a controllable mechanism to adjust the strictness of pseudo-label filtering: a smaller $\alpha$ enhances the utilization of more pseudo-labels, while a larger $\alpha$ suppresses noisy pseudo-labels by raising the filtering threshold. We uniformly set $\alpha = 1$ across all datasets to ensure generalizability without dataset-specific tuning. All experiments are conducted on a single NVIDIA A800 GPU.

## 4.2 Experimental Results

**RQ1:** Does Bi-CoG enhance existing PEFT methods for VLMs in open-world generalization scenarios?

To evaluate the robustness of Bi-CoG in open-world scenarios, we conduct a comprehensive study by integrating it with three representative prompt tuning methods across 14 datasets. Notably, in all experiments related to Bi-CoG, we controlled the random seeds during both the initialization and training phases for different models to ensure diversity in the initial training stages. For each dataset, we report two metrics: HM of base and novel class accuracies, and overall accuracy (Acc.). We set the base-to-novel class ratio to 50:50 and report the mean and standard deviation over three runs. As shown in Table 1, Bi-CoG consistently improves both metrics across all datasets, achieving the highest average performance. Notably, when applied to CoOp, Bi-CoG yields over a 5% improvement in HM. Moreover, Bi-CoG demonstrates particularly strong gains on traditional open-world benchmarks. For example, on low-resolution datasets such as CIFAR-10 and CIFAR-100 [Krizhevsky *et al.*, 2009], it achieves up to an 8% increase in HM, underscoring its effectiveness across datasets with varying resolutions. These results highlight the effectiveness of Bi-CoG and emphasize the importance of integrating unsupervised strategies with pre-trained models in open-world settings.

| Method | CIFAR-10 | | CIFAR-100 | |
|---|---|---|---|---|
| | 4 shots | 25 shots | 4 shots | 25 shots |
| ReMixMatch | $80.90_{\pm 9.64}$ | $94.56_{\pm 0.05}$ | $55.72_{\pm 2.06}$ | $\mathbf{72.57}_{\pm 0.31}$ |
| FixMatch (RA) | $86.19_{\pm 3.37}$ | $94.93_{\pm 0.65}$ | $51.15_{\pm 1.75}$ | $71.71_{\pm 0.11}$ |
| FixMatch (CTA) | $88.61_{\pm 3.35}$ | $\mathbf{94.93}_{\pm 0.33}$ | $50.05_{\pm 3.01}$ | $71.36_{\pm 0.24}$ |
| CoOp | $80.60_{\pm 1.79}$ | $85.10_{\pm 0.29}$ | $49.00_{\pm 0.83}$ | $57.43_{\pm 0.25}$ |
| + *Bi-CoG* | $84.90_{\pm 1.06}$ | $86.10_{\pm 0.03}$ | $52.77_{\pm 0.24}$ | $58.05_{\pm 0.05}$ |
| PromptSRC | $86.03_{\pm 0.48}$ | $90.10_{\pm 0.08}$ | $60.33_{\pm 0.17}$ | $66.33_{\pm 0.29}$ |
| + *Bi-CoG* | $\mathbf{89.83}_{\pm 0.35}$ | $91.07_{\pm 0.42}$ | $\mathbf{60.56}_{\pm 0.79}$ | $66.84_{\pm 0.67}$ |

Table 2: Performance on low-resolution SSL datasets.

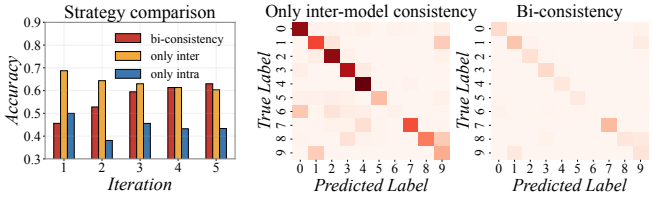
**RQ2:** Can Bi-CoG deliver competitive performance compared to conventional SSL approaches?

As shown in Table 2, SSL algorithms exhibit a significant advantage over baseline prompt tuning methods on SSL datasets. This is largely due to the lower resolution of these datasets, which makes it more difficult for VLMs to effectively capture visual content [Krizhevsky *et al.*, 2009; Lv *et al.*, 2025b]. Notably, this advantage also comes at the cost of substantial computational overhead: conventional SSL methods typically require over 10 hours of training for the CIFAR dataset, whereas Bi-CoG achieves comparable or superior performance with only 1 hour of training under the same hardware configuration. This limitation of high computational cost, coupled with the sensitivity to the number of shots (increasing the number of shots from 4 to 25 yields much larger performance gains for SSL algorithms compared to VLM-based prompt tuning), highlights the practical challenges of traditional SSL methods. In contrast, Bi-CoG provides the model with more reliable training examples via adaptive pseudo-labeling, enabling prompt tuning methods to achieve SOTA performance under the challenging 4-shot setting while maintaining remarkable computational efficiency.

These results demonstrate that Bi-CoG can effectively leverage unlabeled data to capture richer visual representations even in low-resolution and data-scarce scenarios, while also significantly reducing training time compared to SSL methods, making it more practical and scalable.

**RQ3:** How does Bi-CoG compare to other methods that jointly leverage pre-trained models and unlabeled data?

To further validate the effectiveness of Bi-CoG, we conducted comparative experiments against two SOTA methods: GRIP [Menghini *et al.*, 2023] and CPL [Zhang *et al.*, 2024]. Specifically, GRIP defines a fixed epoch-dependent expression for selecting the number of pseudo-labels, while CPL relies on manually set confidence thresholds to filter pseudo-labels iteratively. Following the experimental protocol of GRIP, we adopted the same base-to-novel class ratio (62:38) and used CoOp as the underlying prompt tuning framework for Bi-CoG and all compared methods. This ensures a fair comparison by eliminating discrepancies from different base models or tuning strategies. As shown in Table 3, Bi-CoG achieves the best performance under both SSL




(a) DTD (b) EuroSAT

Figure 3: (a) Pseudo-label accuracy on DTD under different pseudo-labeling strategies; (b) Distribution of pseudo-labels on EuroSAT before and after applying the intra-model consistency.

| Method | Domain Data | SSL | Open-World |
|---|---|---|---|
| CLIP | Zero-shot | 72.08 | 77.80 |
| FPL | Few-shot + Unlabeled | 75.96 | 80.97 |
| IFPL | | 78.68 | 82.08 |
| GRIP | | 83.60 | 86.26 |
| CPL | | 89.66 | 87.35 |
| Bi-CoG | | **96.45** | **88.01** |
| Δ | | +6.79 | +0.66 |

Table 3: Performance comparison on the Flowers102 dataset.

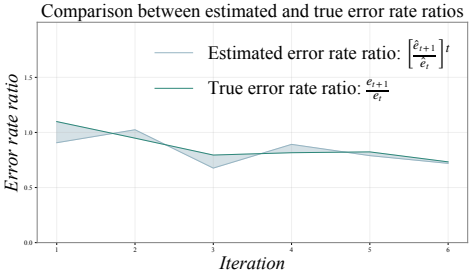



Figure 4: Comparison between the estimated and true error rate ratios. The scaling factor is fixed at $\alpha = 1$ throughout training.

| Method | Open-World | |
|---|---|---|
| | HM | Acc. |
| CoOp + Bi-CoG | 77.79 | 71.03 |
| - *w/o. Inter-model consistency* | 76.14 | 68.83 |
| - *w/o. Intra-model consistency* | 75.85 | 69.08 |
| - *w/o. Error-aware filtering* | 76.92 | 70.92 |

Table 4: Ablation study of Bi-CoG's pseudo-labeling strategy.

and open-world settings, outperforming CPL by up to 5.9 percentage points. Notably, this superiority stems from Bi-CoG's adaptive pseudo-label selection strategy during iterations: unlike GRIP's epoch-dependent rule or CPL's manual threshold tuning, Bi-CoG adjusts pseudo-label selection based on model consistency and estimated error rate, ensuring high-quality supervision signals. More importantly, Bi-CoG dispenses with the requirement for manually tuning hyperparameters, effectively reducing dependence on human expertise. In summary, Bi-CoG represents one of the most effective integrations of pretrain-finetuning and semi-supervised learning methods to date.

### 4.3 Ablation Study

To answer **RQ4**, we first examine the impact of the bi-consistency-guided pseudo-labeling strategy on both pseudo-label accuracy and the distribution of biased samples. As shown in Figure 3.a, removing the inter-model consistency module results in a notable decline in pseudo-label accuracy, highlighting the importance of ensemble-based labeling for generating high-quality supervision. Notably, although the bi-consistency strategy exhibits relatively lower pseudo-label accuracy in the early stages of training, its accuracy steadily improves as training progresses. In contrast, other single-strategy pseudo-labeling methods show either a decline or fluctuation in accuracy, demonstrating the stability and reliability of bi-consistency-guided pseudo-labeling. Furthermore, as illustrated in Figure 3.b, incorporating intra-model consistency helps smooth the class-wise distribution of pseudo-labels, effectively mitigating the bias inherited from pre-training. These results collectively confirm that the bi-consistency-guided strategy produces pseudo-labels that are both more accurate and less biased—two critical factors for stable and effective self-training.

To further validate the theoretical foundation of our error-aware mechanism, we analyze the evolution of the actual error rate ratio on pseudo-labeled samples and its estimation from labeled data on StanfordCars [Krause *et al.*, 2013] under the open-world generalization setting. As illustrated in Figure 4, the condition in Equation 6 holds reliably during training, supporting the practical validity of our theoretical derivation and confirming the soundness of Bi-CoG.

Finally, we conduct a comprehensive ablation study on key components of Bi-CoG, using CoOp as the base model and reporting the average performance across 14 datasets. As shown in Table 4, any component of our three-stage pseudo-labeling strategy, which integrates bi-consistency and error-aware filtering, plays a critical role in improving model performance. This highlights the deeply coupled and effective design of the pseudo-labeling mechanism in Bi-CoG.

## 5 Conclusion

This paper investigates the integration of pre-trained VLMs with semi-supervised learning for downstream tasks with sparse labeled data. We identify critical issues in existing approaches, including the trade-off between label quality and model bias, as well as the limited adaptability arising from the reliance on manually set fixed thresholds. To address these challenges, we propose Bi-CoG, a plug-and-play self-training method that incorporates a deeply coupled pseudo-labeling strategy with error-aware dynamic filtering capabilities. Theoretical analysis and experimental results across 14 datasets in both open-world generalization and semi-supervised settings demonstrate that Bi-CoG can be seamlessly integrated with any pretrain-finetuning methodologies, consistently achieving performance improvements.

## Acknowledgments

This research was supported by the Key Program of Jiangsu Science Foundation (BK20243012), the National Science Foundation of China (62306133), and the "111 Center" (No. B26023).

## References

[Angluin and Laird, 1988] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[Chen *et al.*, 2022] Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debiased self-training for semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2022.

[Gan and Wei, 2024] Kai Gan and Tong Wei. Erasing the bias: Fine-tuning foundation models for semi-supervised learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 14453–14470, 2024.

[Gao *et al.*, 2024] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[Guo and Li, 2022] Lan-Zhe Guo and Yu-Feng Li. Class-imbalanced semi-supervised learning with adaptive thresholding. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8082–8094, 2022.

[Guo *et al.*, 2020] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3897–3906, 2020.

[Guo *et al.*, 2025] Lan-Zhe Guo, Lin-Han Jia, Jie-Jing Shao, et al. Robust semi-supervised learning in open environments. *Frontiers of Computer Science*, 19:198345, 2025.

[Gupte *et al.*, 2024] Sanket Rajan Gupte, Josiah Aklilu, Jeffrey J Nirschl, and Serena Yeung-Levy. Revisiting active learning in the era of vision foundation models. *Transactions on Machine Learning Research*, 2024.

[Huang *et al.*, 2022] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916, 2021.

[Khattak *et al.*, 2023a] 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.

[Khattak *et al.*, 2023b] 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.

[Krause *et al.*, 2013] 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.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[Lee and others, 2013] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013.

[Li *et al.*, 2021] Yu-Feng Li, Lan-Zhe Guo, and Zhi-Hua Zhou. Towards safe weakly supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):334–346, 2021.

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

[Lv *et al.*, 2025a] Song-Lin Lv, Yu-Yang Chen, Zhi Zhou, Ming Yang, and Lan-Zhe Guo. BMIP: Bi-directional modality interaction prompt learning for VLM. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 5887–5895, 2025.

[Lv *et al.*, 2025b] Song-Lin Lv, Rui Zhu, Tong Wei, Yu-Feng Li, and Lan-Zhe Guo. Unlabeled data vs. pre-trained knowledge: Rethinking ssl in the era of large models. *arXiv preprint arXiv:2505.13317*, 2025.

[Menghini *et al.*, 2023] Cristina Menghini, Andrew Delworth, and Stephen Bach. Enhancing CLIP with CLIP: Exploring pseudolabeling for limited-label prompt tuning. In *Advances in Neural Information Processing Systems*, 2023.

[Oh *et al.*, 2024] Changdae Oh, Hyesu Lim, Mijoo Kim, Dongyoon Han, Sangdoo Yun, Jaegul Choo, Alexander Hauptmann, Zhi-Qi Cheng, and Kyungwoo Song. Towards calibrated robust fine-tuning of vision-language models. In *Advances in Neural Information Processing Systems*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th*

*International Conference on Machine Learning*, pages 8748–8763, 2021.

[Shao *et al.*, 2026] Jie-Jing Shao, Bo-Wen Zhang, Xiao-Wen Yang, Baizhi Chen, Siyu Han, Jinghao Pang, Wen-Da Wei, Guohao Cai, Zhenhua Dong, Lan-Zhe Guo, and Yu-Feng Li. ChinaTravel: An open-ended travel planning benchmark with compositional constraint validation for language agents. In *International Conference on Learning Representations*, 2026.

[Tan *et al.*, 2025] Hao-Zhe Tan, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. Vision-language model selection and reuse for downstream adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 58726–58738, 2025.

[Wang and Russakovsky, 2023] Angelina Wang and Olga Russakovsky. Overwriting pretrained bias with finetuning data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3957–3968, 2023.

[Wang *et al.*, 2022] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14647–14657, 2022.

[Yang *et al.*, 2023] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2023.

[Yang *et al.*, 2025] Xiao-Wen Yang, Jie-Jing Shao, Lan-Zhe Guo, Bo-Wen Zhang, Zhi Zhou, Lin-Han Jia, Wang-Zhou Dai, and Yu-Feng Li. Neuro-symbolic artificial intelligence: Towards improving the reasoning abilities of large language models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 10770–10778, 2025.

[Zhang *et al.*, 2022] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *Proceedings of the European Conference on Computer Vision*, pages 493–510, 2022.

[Zhang *et al.*, 2024] Jiahan Zhang, Qi Wei, Feng Liu, and Lei Feng. Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data. In *Proceedings of the 41st International Conference on Machine Learning*, pages 60004–60020, 2024.

[Zhou and Li, 2005] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541, 2005.

[Zhou *et al.*, 2022a] 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.

[Zhou *et al.*, 2022b] 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.

[Zhou *et al.*, 2024] Zhi Zhou, Ming Yang, Jiang-Xin Shi, Lan-Zhe Guo, and Yu-Feng Li. DeCoOp: Robust prompt tuning with out-of-distribution detection. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[Zhu *et al.*, 2022] Zhaowei Zhu, Tianyi Luo, and Yang Liu. The rich get richer: Disparate impact of semi-supervised learning. In *International Conference on Learning Representations*, 2022.