Leveraging Self Weak-supervision for Improved VLM Performance

Shuvendu Roy, Ali Etemad
Queen's University, Canada
{shuvendu.roy, ali.etemad}@queensu.ca

Abstract

In this work, we present SelfPrompt, a novel semi-supervised prompt-tuning approach for tuning vision-language models (VLMs) in a semi-supervised learning setup. Existing methods for tuning VLMs in semi-supervised setup struggle with the efficient use of the limited label set budget, the accumulation of noisy pseudo-labels and proper utilization of the unlabelled data. SelfPrompt addresses these challenges by introducing (a) a weakly-supervised sampling technique that selects a diverse and representative labelled set, (b) a cluster-guided pseudo-labelling method that improves pseudo-label accuracy, and (c) a confidence-aware semi-supervised learning module that maximizes the utility of unlabelled data by learning from high- and low-confidence pseudo-labels differently. We conduct extensive evaluations across 13 datasets, significantly surpassing state-of-the-art performance with average improvements of 7.92% in semi-supervised learning using a 2-shot setup. Our detailed ablation studies show the effectiveness of each component.

1 Introduction

Vision-language models (VLMs) [1] pre-trained on large-scale datasets of image-text pairs have shown strong generalization on a wide range of tasks. Nonetheless, prior works [2, 3] have demonstrated that VLMs require fine-tuning on a considerable amount of labelled data to perform well on downstream tasks. Additionally, the size of the foundation model makes fine-tuning in a limited labelled data setting difficult without losing generalization [4]. To reduce the reliance on labelled data, some recent works have explored semi-supervised solutions that utilize auxiliary unlabelled data [5, 6, 7] to improve learning from a limited set of labelled data.

Although prior works that leverage unlabelled data for tuning VLMs show substantial performance gains, we identify several limitations in such approaches. **(a)** First, given a limited budget for the labelled data (few samples per class), existing methods [7, 6] typically select the labelled sample set randomly. However, a randomly selected set of samples may not adequately represent the underlying data distribution, leading to inefficient use of the limited label budget. **(b)** Next, given the unlabelled set, prior works [6, 7] utilize the zero-shot capabilities of pre-trained VLMs to predict pseudo-labels for the unlabelled data to then use as labelled samples. However, pre-trained VLMs do not necessarily possess adequate knowledge of the downstream domain, which could lead to incorrect pseudo-labels. **(c)** Finally, previous works [6, 7] have employed incremental pseudo-labelling, wherein the labelled set is continuously expanded by iteratively adding to the pseudo-label set from the unlabelled set. Nevertheless, as Figure 1 illustrates, this method often results in the accumulation of noisy pseudo-labels, ultimately leading to performance degradation.

To solve the above-mentioned problems, we propose **SelfPrompt**, a new prompt tuning approach that uses weak supervision by the pre-trained VLM itself to fine-tune the model with a confidence-aware semi-supervised learning approach. SelfPrompt comprises three components. **(a) A weakly-supervised labelled set sampling module:** To select the most representative set of samples for the labelled set, we propose a novel sampling technique. First, the VLM's predictions are used as a source of weak supervision to filter out both the most and least confident samples from the unlabelled set. This is followed by a clustering-based selection technique that identifies a diverse set of samples from the remaining unlabelled data for labelling. **(b) Cluster-guided pseudo-labelling:** To address the second problem, we propose a cluster-guided pseudo-labelling approach that leverages the clusters formed in the labelled set sampling module, we select samples that are near the centroids of the above-mentioned clusters to which we assign the corresponding class labels. **(c) Confidence-aware semi-supervised learning:** To make the best use of the unlabelled data, we propose a confidence-aware semi-supervised module. This hybrid approach leverages high-confidence pseudo-labels in a fully supervised learning setting, while learning from low-confidence samples in a weakly-supervised manner.

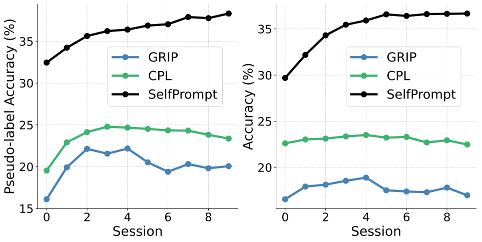

Figure 1: (left) Pseudo-label accuracy and (right) Test accuracy over training sessions.

To evaluate the proposed solution, we follow [6, 7] to perform semi-supervised learning on a 2-shot labelled set setting with the remaining samples of the corresponding dataset used as the unlabelled set. While previous works report these results on six datasets, we perform the evaluation of our solution (and the previous methods) on **13 datasets**. Our evaluation shows that SelfPrompt outperforms prior works by considerable margins, in both textual prompt tuning and visual prompt tuning setups. Specifically, it outperforms prior SOTA by up to 15.05%, with an average improvement of 7.92%. Additionally, we show generalization by reducing the size of the labelled set to just one sample-per-class, while outperforming prior methods by an average of 11.78%. Finally, we present extensive ablation and sensitivity studies on different components of our proposed method.

2 Related Works

Prompt tuning is a parameter-efficient technique for adapting foundation models to downstream tasks by learning soft prompts (textual [8, 9] or visual [10, 11]) from limited labelled data [8, 11]. Text-based prompt tuning [9] involves optimizing learnable prompt vectors, which are embedded into the input sentence tokens fed into the sentence encoder. Recently, a new stream of research has focused on semi-supervised tuning that leverages unlabelled data alongside a small set of labelled data to enhance downstream task performance. The core idea behind these methods is pseudo-labelling [12, 13], where the model predicts labels for unlabelled samples and uses these pseudo-label to learn from the unlabelled data. For example, GRIP [6] utilizes CLIP's zero-shot capabilities to generate pseudo-labels for unlabelled data and select the most confident samples to serve as labelled data. However, this approach introduces a considerable amount of wrong pseudo-labels due to the inherent miscalibration [14] and imbalanced predictions [15] issues of the pre-trained VLM. To address these issues, CPL [7] proposes to generate refined candidate pseudo-labels through intra- and inter-instance label selection, using a confidence score matrix to improve label accuracy and class balance during fine-tuning. Both GRIP and CPL adopt an iterative process, where the model is used to continuously refine and select additional samples from the unlabelled set.

3 Method

Let θ be a pre-trained image encoder and ϕ be a text encoder of a pre-trained VLM. For a given input image x, the VLM predicts the output probability distribution over C classes as:

$$p(y|x) = \frac{\exp(\text{sim}(z, w_y)/\tau)}{\sum_{k=1}^{C} \exp(\text{sim}(z, w_k)/\tau)}, \tag{1}$$

where $z = \theta(x)$ is the image embedding and w_k is the class-embedding of class k, generated using a prompt template as $w_k = \phi(\text{`a photo of a [category]}_k\text{'})$. A recent class of solutions proposes the use of semi-supervised learning for tuning VLMs by leveraging a large unlabelled set along with a small labelled set. Despite recent progress in semi-supervised prompt tuning for VLMs, we have identified three key open challenges in this area, including the under-utilization of the labelling budget, the negative impact of miscalibrated pseudo-labels, and the declining quality of pseudo-labelling as the number of samples increases.

In light of the challenges above, we propose SelfPrompt, a novel semi-supervised prompt tuning method for VLMs that introduces three novel components: a weakly-supervised sampling module, cluster-guided pseudo-labelling, and confidence-aware semi-supervised learning.

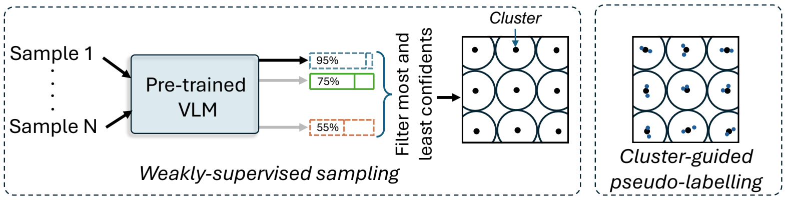
Sample 1
Sample N
Pre-trained VLM
95%
75%
55%
Filter most and least confidents
Cluster
Weakly-supervised sampling
Cluster-guided pseudo-labelling

Figure 2: (left) Visual illustration of the weakly-supervised sampling. (right) Cluster-guided pseudo-labelling.

Weakly-supervised sampling. To overcome the limitations of random selection, we introduce a weakly supervised sampling module that selects the most diverse and representative N samples from the unlabelled set. This module operates through a two-step protocol:

Step 1: Filtering with weak supervision. We leverage the zero-shot predictions of the pre-trained VLM as weak supervision to filter the unlabelled set U. Specifically, we remove samples with both the highest and lowest confidence predictions by the VLM. Highly confident samples offer minimal information gain, as the model is already certain of their classification. Conversely, low-confidence samples are likely to be outliers or noisy data points that can negatively impact model generalization, especially in few-shot learning scenarios where training data is scarce. For each unlabelled sample i, we generate a probability distribution over the output classes with the pre-trained VLM using Eq. 1 as: $\mathbf{p}_i = [p_i^1, p_i^2, \cdots, p_i^C]$. We define the confidence score for each sample as the maximum probability value over the classes: $c_i = \max_{1 \leq c \leq C} p_{ic} = \max\{p_i^1, p_i^2, \cdots, p_i^C\}$. We then sort the samples in descending order of confidence: $\mathcal{D}_{\text{sorted}} = \{x_{(1)}, x_{(2)}, \ldots, x_{(N)}\}$, where $x_{(i)}$ is the sample with the i-th highest confidence score, satisfying: $c_{(1)} \geq c_{(2)} \geq \ldots \geq c_{(N)}$. Next, we divide the sorted samples into q quantiles, $\{Q_1, Q_2, \cdots Q_q\}$, and remove the first and last quantiles, corresponding to the most and least confident samples. Finally, the filtered unlabelled dataset after the first step can be represented as $\mathcal{D}_{\text{filtered}} = \bigcup_{k=2}^{q-1} \mathcal{Q}_k$.

Step 2: Diversity Sampling. Next, we select N diverse samples from the filtered dataset $\mathcal{D}_{\text{filtered}}$, with a cluster-based sampling technique. First, we obtain the representations for each sample using a pre-trained vision encoder θ as $\mathbf{z}_i = \theta(x_i) \in \mathbb{R}^d$, where $\mathbf{z}_i$ is the d-dimensional embedding of sample x_i. We then apply k-means clustering to group the samples into N clusters $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$, such that each cluster contains semantically similar samples, while different clusters have diverse semantics. We select the sample closest to the cluster center from each cluster by x_j^*. Finally, our labelled set is formed by gathering the labels of the selected samples, $X_L = \{(x_1^*, y_1), (x_2^*, y_2), \cdots, (x_N^*, y_N)\}$. The proposed module is illustrated in Figure 2 (left).

Cluster-guided pseudo-labelling. To improve the pseudo-label quality, especially at the beginning of the training, we propose a novel clustering-guided pseudo-labelling approach that does not rely on the VLM to generate the pseudo-labels. Instead, our proposed solution leverages the clusters ($\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$) formed during the weakly supervised sampling step. Since the clusters are formed based on embedding similarity, samples under the same cluster have similar semantics. Especially the samples close to the cluster centres (also close to the selected labelled sample) are likely to belong to the same class as the sample at the cluster's center. Implicating this realization, we select additional p samples from each cluster and label them with the label of the cluster center. Specifically, for each cluster $\mathcal{C}_j$, we pick the p samples closest the cluster centers to form a pseudo-label set $\mathcal{P}_j = \{x_j^1, x_j^2, \ldots, x_j^p\}$, where $\mathcal{P}_j$ is the pseudo-label set for cluster $\mathcal{C}_j$, and x_j^k is the k-th closest sample to x_j^*. Finally, each sample in P_j is assigned to the label of the cluster center of $\mathcal{C}_j$ to form our pseudo-label set $\mathcal{X}_p = \{(x_{j1}, y_j), (x_{j2}, y_j), \ldots, (x_{Np}, y_N)\}$. Our cluster-guided pseudo-labelling technique is illustrated in Figure 2 (right).

Confidence-aware semi-supervised learning. To make the best use of the unlabelled data, we propose a confidence-aware semi-supervised module that learns from the high-confident samples in a supervised learning setup, while learning from the low-confident samples in a weakly-supervised setting. Specifically, we first predict the output probability distribution for each sample in the unlabelled set U as $\mathbf{p}_i = f(x_i) \in \mathbb{R}^C$. Then we incorporate the t (defined as $\tau \times M$) most confident samples-per-class into our pseudo-label set as $\mathcal{X}^+ = X_P \cup \left(\bigcup_{c=1}^{C} \text{top}_t(\{x_i | \arg\max(\mathbf{p}_i) = c\})\right)$, where τ is a hyper-parameter. We learn from the remaining relatively low-confident samples in a weakly-supervised setting. Specifically, we follow CPL [7], and gather the top-k predictions per sample to form a weakly-labelled set $\mathcal{X}_{weak} = \{(x_i, s_i) | x_i \notin \mathcal{X}^+\}$, where s_i is a one-hot vector containing $s_{i,c} = 1$ if class c is among the top predictions for sample i. Finally, we learn from the

Table 1: Comparison results of top-1 test accuracy (%) on 13 benchmarks on the **semi-supervised** learning with **textual prompt** strategy.

Methods	Average	Flowers102	RESISC45	DTD	CUB	EuroSAT	FGVCAircraft
Zero-shot CLIP	55.17	$63.67_{0.00}$	$54.48_{0.00}$	$43.24_{0.00}$	$51.82_{0.00}$	$32.88_{0.00}$	$17.58_{0.00}$
CoOp	62.28	$75.96_{0.74}$	$68.13_{0.55}$	$37.10_{5.45}$	$55.29_{0.59}$	$62.05_{1.64}$	$20.02_{0.77}$
GRIP	67.40	$83.60_{0.48}$	$74.11_{0.68}$	$56.07_{0.79}$	$56.65_{0.33}$	$58.66_{2.64}$	$16.98_{0.20}$
CPL	71.41	$89.66_{0.36}$	$80.98_{0.11}$	$61.21_{0.56}$	$58.53_{0.24}$	$77.51_{0.80}$	$22.48_{0.63}$
SelfPrompt	**79.33**	$\mathbf{93.04_{0.33}}$	$\mathbf{85.58_{0.18}}$	$\mathbf{72.18_{0.78}}$	$\mathbf{68.84_{0.16}}$	$\mathbf{87.49_{0.12}}$	$\mathbf{36.71_{0.70}}$
Δ	↑ 7.92	↑ 3.38	↑ 4.60	↑ 10.97	↑ 12.31	↑ 9.98	↑ 14.23

	Caltech101	MNIST	Food101	StanfordCars	OxfordPets	SUN397	UCF101
Zero-shot CLIP	$82.01_{0.00}$	$25.10_{0.00}$	$78.81_{0.00}$	$60.29_{0.00}$	$84.32_{0.00}$	$62.54_{0.00}$	$60.42_{0.00}$
CoOp	$84.69_{1.43}$	$58.22_{1.98}$	$76.23_{1.45}$	$58.23_{2.45}$	$82.34_{1.44}$	$62.19_{1.78}$	$69.19_{1.03}$
GRIP	$85.99_{1.06}$	$71.78_{2.59}$	$80.89_{1.14}$	$62.83_{1.42}$	$89.40_{0.33}$	$67.34_{0.98}$	$71.94_{0.95}$
CPL	$92.87_{1.14}$	$75.18_{4.40}$	$79.38_{1.05}$	$61.93_{1.30}$	$87.79_{1.31}$	$66.98_{0.65}$	$73.88_{1.32}$
SelfPrompt	$\mathbf{94.10_{0.92}}$	$\mathbf{90.23_{0.36}}$	$\mathbf{82.19_{0.17}}$	$\mathbf{75.21_{0.33}}$	$\mathbf{89.86_{0.48}}$	$\mathbf{74.77_{0.18}}$	$\mathbf{81.07_{0.44}}$
Δ	↑ 1.23	↑ 15.05	↑ 2.81	↑ 13.28	↑ 2.07	↑ 7.79	↑ 7.19

labelled set $\mathcal{X}_L$, pseudo-labeled set $\mathcal{X}^+$, and weakly labelled set $\mathcal{X}_{weak}$, together as follow:

$$\mathcal{L}_{final} = \frac{1}{|\mathcal{X}_L|} \sum_{(x,y)\in\mathcal{X}_{\mathcal{L}}} \ell(f(x), y) + \frac{1}{|\mathcal{X}+|} \sum_{(x,y)\in\mathcal{X}+} \ell(f(x), y) + \frac{\lambda}{|\mathcal{L}_{weak}|} \sum_{(x,s)\in\mathcal{L}weak} \ell_w(f(x), s). \quad (2)$$

Here, ℓ is the cross-entropy loss and ℓ_w is a partial label learning loss from CPL [7].

4 Experiments

Implementation details. Following [7] and [6] we adopt a CLIP ViT-B/32 [1] as the pre-trained backbone of our model. All experiments are conducted in a 2-shot setup, with ten sessions of iterative pseudo-labelling (50 epochs per session). The model is optimized using SGD with a learning rate of 0.02 and a batch size of 64. Results are reported as the average accuracy over three runs.

Results. First, we present our results on semi-supervised learning. The results of this experiment are presented in Table 1, where we observe that SelfPrompt shows large and consistent improvements over prior works across the 13 datasets. On average, SelfPrompt achieves an accuracy of 79.33% with just two labelled samples per class, which is a 7.92% improvement over the previous SOTA CPL and a 12.04% improvement over GRIP. Notably, SelfPrompt shows up to 15.05% improvement over the previous SOTA on individual datasets. More importantly, SelfPrompt shows higher improvements on datasets with lower zero-shot (VLM) accuracies (e.g., FGVCAircraft and MNIST).

Table 2: Ablation study. W.S.S., C.G.P., and C.A.SSL correspond to weakly-sup sampling, cluster-guided pseudo-labelling, and confidence-aware SSL.

W.S.S.	C.G.P.	C.A.SSL	Accuracy
✓	✓	✓	79.33
✗	✓	✓	76.12
✓	✗	✓	74.39
✓	✓	✗	78.01
✗	✗	✓	73.49
✗	✓	✗	75.67
✓	✗	✗	73.08
✗	✗	✗	71.41

We present an ablation study on our proposed method in Table 2. Here, W.S.S., C.G.P., and C.A.SSL correspond to the three modules of our proposed solution, namely, weakly-supervised sampling, cluster-guides pseudo-labelling, and confidence-aware semi-supervised learning. The results are reported as the average accuracy over all datasets. Here, including all components has an average accuracy of 79.33%, while removing all components (the baseline) has an accuracy of 71.41%. As we find from this table, cluster-guides pseudo-labelling has a high impact on the overall performance of our proposed solution, removing which results in a 4.94% drop in the performance. This is also evident from the fact including only this component shows a 4.26% improvement over the baseline. Next, removing weakly-supervised sampling shows a 3.21% drop in performance, while adding only these components shows a 1.61% gain in performance. This is due to the fact that without weakly supervised sampling, cluster-guides pseudo-labelling with the randomly selected samples do not show the best performance. Finally, removing confidence-aware semi-supervised learning also shows a 1.32% drop in performance.

5 Conclusion

In this paper, we propose SelfPrompt, a novel semi-supervised tuning approach for vision-language models that addresses three key limitations of prior methods: under-utilization of limited labeled data, miscalibrated pseudo-labeling, and the negative impact of noisy labels. SelfPrompt outperforms previous works by an average of 7.92% in semi-supervised learning and 4.9% in base-to-novel generalization across 13 datasets. Notably, it achieves strong generalization with just one labeled sample per class, demonstrating its effectiveness in real-world tasks with minimal labeled data.

References

[1] 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 *ICML*, pages 8748–8763, 2021. 1, 4

[2] 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. 1

[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1

[4] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *ICLR*, 2024. 1

[5] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 1

[6] Cristina Menghini, Andrew Delworth, and Stephen Bach. Enhancing clip with clip: Exploring pseudolabeling for limited-label prompt tuning. *NeurIPS*, 36:60984–61007, 2023. 1, 2, 4

[7] Jiahan Zhang, Qi Wei, Feng Liu, and Lei Feng. Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data. In *ICML*, 2024. 1, 2, 3, 4

[8] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[9] 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. 2

[10] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 2

[11] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2

[12] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, pages 596–608, 2020. 2

[13] 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. 2

[14] Will LeVine, Benjamin Pikus, Pranav Raja, and Fernando Amat Gil. Enabling calibration in the zero-shot inference of large vision-language models. *arXiv preprint arXiv:2303.12748*, 2023. 2

[15] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In *CVPR*, pages 14647–14657, 2022. 2