This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Beyond Words: Augmenting Discriminative Richness via Diffusions in Unsupervised Prompt Learning

Hairui Ren[1] Fan Tang[2] He Zhao[3] Zixuan Wang[1] Dandan Guo[1*] Yi Chang[1*]
[1] School of Artificial Intelligence, Jilin University
[2] Institute of Computing Technology, Chinese Academy of Sciences [3] CSIRO's Data61
{renhr22,zixuan24}@mails.jlu.edu.cn,tangfan@ict.ac.cn,he.zhao@ieee.org,{guodandan,yichang}@jlu.edu.cn

Abstract

*Fine-tuning vision-language models (VLMs) with large amounts of unlabeled data has recently garnered significant interest. However, a key challenge remains the lack of high-quality pseudo-labeled data. Current pseudo-labeling strategies often struggle with mismatches between semantic and visual information, leading to sub-optimal performance of unsupervised prompt learning (UPL) methods. In this paper, we introduce a simple yet effective approach called **A**ugmenting **Di**scriminative **R**ichness via Diffusions (AiR), toward learning a richer discriminating way to represent the class comprehensively and thus facilitate classification. Specifically, our approach includes a pseudo-label generation module that leverages high-fidelity synthetic samples to create an auxiliary classifier, which captures richer visual variation, bridging text-image-pair classification to a more robust image-image-pair classification. Additionally, we exploit the diversity of diffusion-based synthetic samples to enhance prompt learning, providing greater information for semantic-visual alignment. Extensive experiments on five public benchmarks, including RESISC45 and Flowers102, and across three learning paradigms-UL, SSL, and TRZSL-demonstrate that AiR achieves substantial and consistent performance improvements over state-of-the-art unsupervised prompt learning methods. Code is available.*

1. Introduction

Large pre-trained vision-language models (VLMs) [12, 24, 32] have achieved impressive performance in open-world visual concept learning without requiring task-specific training; however, they still benefit from adaptation for optimal results. Prompt tuning [7, 13, 25, 26] offers an efficient strategy to enhance VLMs' discriminative accuracy on downstream tasks by learning inputs to the model.

While prompt learning can yield notable improvements, a large amount of labeled data is still required to enhance

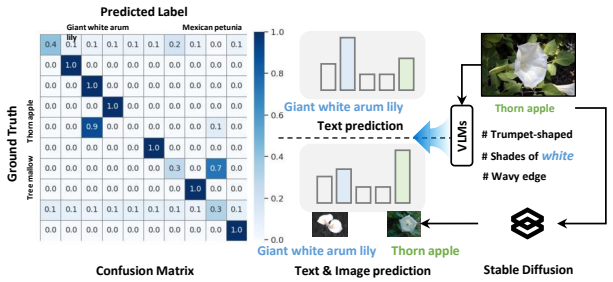

Figure 1. Left: Confusion matrix for ground truth labels $vs.$ text classifier predictions on the Flowers102 dataset, highlighting persistent generation of incorrect pseudo-labels. Right: An example of misclassification in Flowers102, where "Thorn apple" is misidentified as "Giant white arum lily" due to its white petals resembling the latter's semantic information. This misclassification can be alleviated by jointly considering text and image predictions.

VLMs' performance across diverse downstream tasks. This adaptation requirement leads to considerable labeling costs, as labeled data is hard to obtain [14, 30, 35, 36]. To address this, unsupervised prompt learning (UPL) [11, 17, 21, 28] has been introduced, leveraging the zero-shot capabilities of VLMs to generate pseudo labels, enabling effective use of unlabeled data. For pseudo-labeling, traditional methods [11] typically form a text classifier by inputting prompts with class names into the VLMs' text encoder. Then, image embeddings of unlabeled data are matched to these class-based text classifiers, assigning pseudo-labels to images with high-confidence predictions. Some other approaches [21, 33] generate multiple candidate pseudo-labels by adjusting confidence thresholds or optimize the model with different numbers of pseudo-labeled samples at various training stages. While these methods apply distinct strategies for filtering pseudo-labels or refining prompts, most ultimately rely on pseudo-labels produced by text classifiers, making the quality of these pseudo-labels critically important.

In this work, we aim to improve the pseudo label quality

in UPL, which can thus be easily combined with other UPL methods. Here, we first explore conduct a pilot experiment to empirically demonstrate the deficiency of text classifiers.

Specifically, we leverage the text encoder of CLIP [24] to generate pseudo labels on the Flowers102 dataset [22] and subsequently calculated the confusion matrix between the ground-truth labels and predicted class labels. The results, as shown in Fig. 1, reveal that a substantial number of samples are incorrectly predicted, $e.g.$, most "Thorn apple" flowers are misclassified as "Giant white arum lily", "Tree mallow" is misclassified as "Mexican petunia". Why this happened? The discrimination of unlabeled images heavily relies on the precision of text classifiers, which are constructed using the words of prompt with category label names. However, when only using the text of "a photo of a + class name" to describe the image, using "Giant white arum lily" as the class name is even closer in semantics to the given image than "Thorn apple". Because the Thorn apple flower appears to be a shade of white but the word "Thorn apple" does not express that, resulting in the misclassified samples. It motivates us to learn a richer discriminating way to represent the class comprehensively and thus facilitate classification. Since an image has rich information, we consider a novel solution: in addition to the text classifier built on the short text description, can we design another auxiliary classifier based on the most representative images for each category? However, this solution is problematic since we do not know the ground-truth label of each training image in the unsupervised prompt learning task, making it a dead-end loop.

To address the above problem, in this paper, we draw inspiration from the image generation literature. Recent advancements in image generation have made it possible to generate fidelity and diverse images with only text word descriptions, such as Stable Diffusion [27]. It is effective in generating diverse images with richer visual appearance variation while preserving the key semantics. Surprisingly, we find that this text-to-image generation model can be naturally utilized to generate the synthetic representative images for each class in UPL tasks only with the text descriptions on the class names. Therefore, we propose Beyond Words: **A**ugmenting D**i**scriminative **R**ichness via Diffusions, a new framework **AiR** which applies diffusion models to generate synthetic images, which serve as a bridge from text-image-pair to image-image-pair classification, as a way to discriminate the unlabeled images. In our AiR framework, we adopt Stable Diffusion [27] to generate discriminative richness, where we fine-tune the Stable Diffusion model with LoRA [10] by means of pseudo-labeled samples with high confidence to ensure prediction fidelity. Then, we introduce cosine similarity-based filtration to select the most representative synthetic image of each class generated by fine-tuned Diffusion, which is used to build the auxiliary classifier from the view of the image. Combining the predictions of the original text classifier and the auxiliary classifier, we can endow the training images with more accurate pseudo labels, improving the UPL performance. Besides, the diffusion-based synthetic samples can also be utilized to learn the prompt. Notably, our proposed AiR only needs to know the dataset name and class name of downstream tasks, which can be seamlessly integrated into most UPL methods. Experimental results show that AiR achieves a notable improvement in classification accuracy by an average of 2.6% in comparison to the state-of-the-art UPL method. To sum up, our contributions can be included:

- We highlight that relying solely on the text classifier to generate pseudo labels in UPL is inadequate and propose a richer discriminating way to represent the class comprehensively and thus facilitate classification.
- We present a novel approach AiR: **A**ugmenting D**i**scriminative **R**ichness via Diffusions. This method leverages the diversity and fidelity of synthetic samples to build an auxiliary classifier that captures richer variations in visual appearance while preserving essential semantics, bridging text-image-pair classification to image-image-pair classification.
- Our AiR demonstrates substantial improvement across three distinct unsupervised paradigms and five widely-used datasets, significantly outperforming the current state-of-the-art UPL method by an average margin of over 2.6%.

2. Related Work

2.1. Prompt Learning

Prompt learning is a technique that enhances the practical application of large pre-trained models like CLIP [24] and GPT [3, 23] by enabling improved performance across various downstream tasks through tailored text or visual prompts. For instance, CoOp [36] optimizes continuous prompt vectors instead of relying on discrete ones, thereby avoiding the suboptimal performance caused by hand-designed prompts. CoCoOp [35] addresses CoOp's generalization challenges by incorporating Meta-Net, which balances performance between base and novel classes. When only unlabeled data is available, some work is dedicated to unsupervised learning to accomplish prompt learning. UPL [11] using CLIP's [24] zero-shot capabilities, followed by confidence refinement to generate pseudo-labels with higher confidence for unsupervised learning. IFPL and GRIP [21], proposed 'Iterative Refinement of Few-pseudo labels' and 'Grow and Refine Iteratively Pseudolabels' respectively, in terms of how to utilize pseudo-labelled samples. To alleviate the issue of hard pseudo labels, CPL [33], proposes a Candidate Pseudolabel Learning method, which progressively generates refined candidate pseudo la-

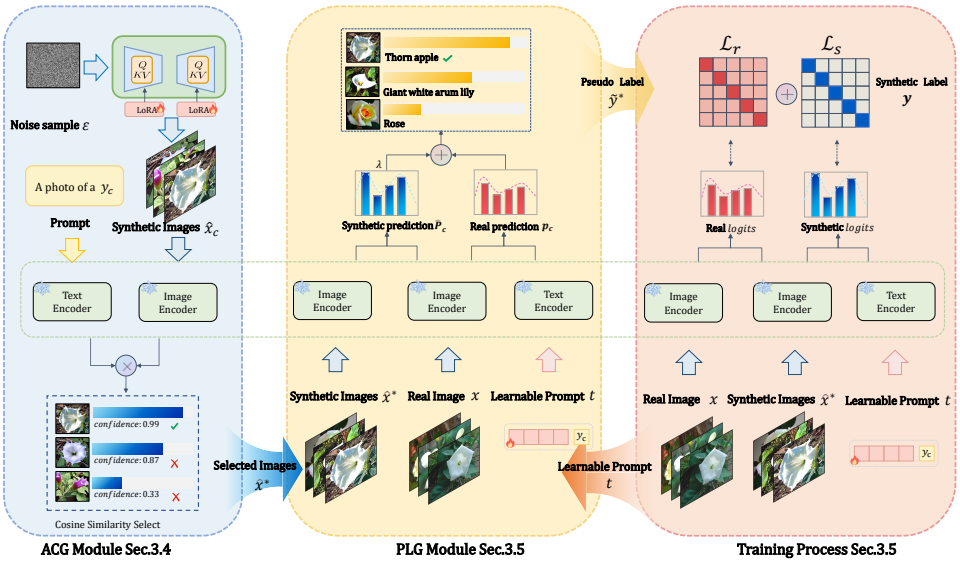

Figure 2. Overview of our proposed AiR, which consists of ACG (Auxiliary Classifier Generation) Module and PLG (Pseudo Label Generation) Module. ACG generates synthetic samples using the LoRA fine-tuned SD model and selects the representative samples by cosine similarity to build auxiliary classifier. PLG generates pseudo-labels by fusing the predictions of the auxiliary classifier built on synthetic images and the text classifier. The overall loss for training prompt consists of the classification loss of real images with pseudo-labels $\mathcal{L}_r$ and classification loss of synthetic images with corresponding categories $\mathcal{L}_s$, where ours can learn visual and textual prompt.

bels by both intra- and inter-instance label selection. Different from these pseudo-labeling [11, 21, 33] approaches, which typically rely on text classifiers to generate pseudo labels. We propose the AiR framework, which uses synthetic images as a bridge to expand text-image-pair classification with image-image-pair classification, further improving the model's discriminative performance from a richer visual perspective.

2.2. Training with Synthetic Data

The diffusion model (DM) has shown significant effectiveness in various synthetic tasks, including image [9, 29], video [2], audio [16], and text [18]. Concurrently, DM has made notable advancements in image manipulation and restoration tasks, such as image editing [1], inpainting [19], and deblurring [31]. In the field of prompt learning, exploration has also been undertaken. For example, DiffTPT [6] generates diverse augmented images with diffusion models while faithfully preserving the key semantics. DataDream [15], a novel few-shot method that adapts Stable Diffusion to generate better in-distribution images for downstream training and is also proven effective in the field of prompt learning. However, most of the aforementioned methods primarily focus on leveraging diffusion models (DM) for data augmentation or generating more realistic images, without employing DM as a bridge to connect the unsupervised text modality with the image modality. Furthermore, the application of DM within unsupervised prompt learning (UPL) remains largely unexplored.

3. Method

3.1. Preliminaries

Revisiting Unsupervised Prompt Tuning. We provide a brief introduction to the prevailing unsupervised prompt learning (UPL) method. In the UPL setting, we have access only to unlabeled data $X_U = \{x_i\}_{i=1}^N$ composed of N instances with target classes name $\{y_1, ..., y_C\}$. Suppose a VLM with learnable prompt parameters t. Besides, in the CLIP zero-shot setting, "a photo of a" is a usually used prompt, denoted as t_{zs}. Each class name $\{y_c\}_{c=1}^C$ is concatenated with a text prompt to formulate a text input "a photo of a [CLS]", which is represented as $w_c = \{t_{zs}, y_c\}$ and can be fed into the text encoder G in VLM to yield a text embedding $g_c = G(w_c)$. Concurrently, given an instance x in X_U, we can embed it with image encoder F of CLIP to produce a visual embedding $f = F(x)$. The prediction probabilities over classes are given as follows:

$$p_c = \frac{\exp(\text{sim}(f, g_c)/\tau)}{\sum_{c'=1}^{C} \exp(\text{sim}(f, g_{c'})/\tau)}, \tag{1}$$

where τ represents the temperature coefficient. Besides, $\text{sim}(\cdot)$ refers to a metric function for measuring the similarity between the input image and the text, which is realized with cosine similarity. Mainstream UPL methods such as FPL [11], GRIP [21] employ CLIP's zero-shot capability to assign a pseudo-label $\tilde{y}$ to each sample x as follows:

$$\tilde{y} = \underset{c}{\text{argmax}} p_c, \tag{2}$$

typically, one can use a top-K pseudo labeling approach, where the top-K most confident examples per class are used as pseudo-labeled data. Then, the prompts t can be optimized via back-propagating the following loss:

$$\min_{t} \mathcal{L}_{ce} = \min_{t} \mathbb{E}_{(x,\tilde{y})} \left[-\log p(\tilde{y}|x) \right]. \tag{3}$$

However, due to the limited expressive power of purely semantic information, the accuracy of pseudo-labels generated by text-only classifiers remains sub-optimal.

3.2. Overview

This work aims to improve the pseudo-labeling accuracy in UPL from a novel view. Specifically, we extend text-image-pair classification to image-image-pair classification by generating synthetic samples for each category, enhancing the classifier's visual perception with the Diffusion Model (DM), which serves as a bridge. The overall workflow of our method can be divided into three steps: ① Finetune DM with LoRA. Considering the need to ensure the diversity of the synthetic samples while making the synthetic samples more consistent with the domain distribution of the current downstream dataset, we use the LoRA strategy to finetune the DM(see Sec. 3.3). ② **ACG** (Auxiliary Classifier Generation) Module: After finetuning with LoRA, we generate synthetic images through the DM with descriptions of the dataset and class name. Then, we obtain the auxiliary classifier with cosine similarity and select a representative sample for each class. (see Sec. 3.4). ③ **PLG** (Pseudo Label Generation) Module: Based on the synthetic image, we expand text-image-pair prediction with image-image-pair prediction, further improving the model's discriminative performance from a richer visual perspective and obtaining a more comprehensive pseudo label (see Sec. 3.5).

3.3. Finetune DM with LoRA

While the DM can generate a diverse range of images through text-to-image generation, these images do not always align with the dataset of the downstream task. This may introduce spurious classes and style mismatches, leading to domain inconsistency and degraded prompt tuning performance. Thus, it is necessary to balance the augmented images' data diversity and domain consistency. The most straightforward approach would be to fine-tune the DM using labeled datasets from downstream tasks; however, this method is computationally expensive, and labeled data is unavailable. To this end, we adopt CLIP's zero-shot generated pseudo-labeled data $\{x, \tilde{y}\}$ in Sec. 3.1 to finetune the DM with LoRA [10], where we employ Stable Diffusion (SD) and the details about the LoRA is deferred to Appendix. The original SD model parameters θ are kept frozen and only the LoRA weights δ are trained with the objective function:

$$\min_{\delta} \mathcal{L}_{lora} = \min_{\delta} \mathbb{E}_{(x,\tilde{y}),\epsilon\sim\mathcal{N}(0,1)}([||\epsilon - \epsilon_{\theta,\delta}(x, t_{data}, \tilde{y})||_2^2], \tag{4}$$

where $\epsilon \sim \mathcal{N}(0,1)$ denotes the sampled noise, $\epsilon_{\theta,\delta}(\cdot)$ denotes the noise generated by SD model. In addition to the training image x and its pseudo-label $\tilde{y}$ generated by zero-shot CLIP, the description of the dataset t_{data} is also the input of the SD model, to ensure that the generation of samples is more in line with the current dataset. Taking the Flowers102 dataset as the example, t_{data} is "Flowers:" and $\tilde{y}$ may be "Thorn apple" or any class else. In this way, we ensure both the diversity of the synthesized samples and domain consistency to the current dataset domain. For subsequent synthetic images, we used the LoRA-updated SD model for generation.

3.4. ACG Module

After fine-tuning with LoRA, SD is presented to generate diverse and domain-consistency data. For each class in the UPL task, we can generate synthetic data with the SD model $S(\cdot)$ as:

$$\hat{X}_c^M = S(\{t_{data}, y_c\}, \epsilon), \tag{5}$$

where $\hat{X}_c^M = \{\hat{x}_c^i\}_{i=1}^M$ represents the generated synthetic images of class y_c and M denotes the number of images generated for each class. Considering the synthetic image needs to be used as the auxiliary classifier, it is essential to select the representative sample from $\hat{X}_c^M$ that best aligns with the semantic information of the current category. Following [6], we compute the cosine similarity between the zero-shot CLIP text embedding g_c and each synthetic image in $\hat{X}_c^M$. Using Eq. 1, we then select the synthetic sample with the highest confidence p_c for each class c and achieve the following synthetic dataset:

$$\hat{X}^* = \{\hat{x}_1^*, \hat{x}_2^*, ..., \hat{x}_C^*\}. \tag{6}$$

We now build the bridge between the text class name and the representative synthetic dataset. Here, $\hat{x}_c^*$ is the most representative synthetic image for the c-th class and it contains the ample visual perceptual discrimination information, exactly what the textual class name used in zero-shot CLIP lacks. Therefore, it is natural to extract the visual embedding of $\hat{X}^*$ with the vision encoder F and build the auxiliary classifier based on the visual prototypes.

3.5. PLG Module & Learning with Synthetic Image

Through the above auxiliary classifier, we can correct the original prediction probabilities in Eq. 1 by the zero-shot CLIP as:

$$\begin{aligned} p_c^* &= p_c + \lambda * \hat{p}_c, \\ \hat{p}_c &= \frac{\exp(\text{sim}(f, F(\hat{x}_c^*))/\tau)}{\sum_{c'=1}^{C} \exp(\text{sim}(f, F(\hat{x}_{c'}^*))/\tau)}, \end{aligned} \tag{7}$$

where $\lambda \geq 0$ serves as a hyper-parameter of these two type predictions. By combining the two predictions, the model's discrimination of the current image contains both semantic and visual perceptual information. Depending on the discrimination result in Eq. 7, we are able to obtain a more comprehensive pseudo label $\tilde{y}^*$ for each sample x as $\tilde{y}^* = \underset{c}{\text{argmax}} p_c^*$, then, we can use these pseudo-labeled data to optimize the prompt t, and the training loss can be expressed as:

$$\min_t \mathcal{L}_r = \min_t \mathcal{L}_{ce}(x, \tilde{y}^*). \tag{8}$$

Considering that the synthetic samples generated by stable diffusion in Eq. 5 also extend the diversity information of the current category while being labeled counterparts, we participate in the synthetic samples in the training process as well as:

$$\min_t \mathcal{L}_s = \min_t \mathcal{L}_{ce}(\hat{x}, y). \tag{9}$$

In conclusion, the training objective of our method can be formally expressed as:

$$\min_t \mathcal{L} = \min_t (\mathcal{L}_r + \beta * \mathcal{L}_s), \tag{10}$$

where $\beta \geq 0$ represents the coefficient of equilibrium between the two loss functions. After each iteration, the optimized prompt t is sent back to the PLG module and processed through Eq. 7 to generate a more accurate pseudo label. This iterative process facilitates the continuous refinement and optimization of pseudolabels. Thus, we not only obtain high-quality pseudo-labeled samples that are classified from semantic and visual perspectives with richer visual perception information but also a diversity of high-quality synthetic samples to aid in training, thus improving the unsupervised prompt learning process.

3.6. The Overall Training Process

In this subsection, we outline the overall training process of our method, AiR. To balance data diversity and domain consistency in the augmented images, we begin by fine-tuning SD with LoRA. Next, synthetic images $\hat{X}$ are generated through SD, and we select the sample best aligned with the semantic information based on cosine similarity. The chosen synthetic images $\hat{X}^*$ then act as auxiliary classifiers, in combination with text classifiers, to assign pseudo-labels $\tilde{y}^*$ to the unlabeled samples. Finally, the pseudo-labeled data $(x, \tilde{y}^*)$ and synthetic data $(\hat{x}, y)$ serve as training samples to calculate the loss and iteratively optimize the prompt. The algorithm of AiR is summarized in the Appendix.

4. Experiments

4.1. Benchmark Setting

Datasets. We conduct a comprehensive evaluation of our method on five diverse classification datasets: DTD [5], EuroSAT [8], FGVC [20], Flowers102 [22], and RESISC45 [4]. These datasets span a wide array of topics, including general object, scene, and fine-grained classification, along with specialized tasks such as texture recognition and satellite imagery analysis.

Learning Paradigms. To comprehensively evaluate the effectiveness of our method, we investigate three common scenarios involving unlabeled data: Unsupervised Learning (UL), Semi-Supervised Learning (SSL), and Transductive Zero-Shot Learning (TRZSL). Detailed explanations of each paradigm and our approach to labeled data are provided in the Appendix.

Baselines. To evaluate the effectiveness of the training strategies outlined in Sec. 3, we compare ours against CLIP's zero-shot performance using default prompts such as 'a photo of a [CLASS]'. We also benchmark against standard supervised prompt-tuning methods that use only labeled data: CoOp [36] for textual prompts and VPT [13] for visual prompts. Additionally, we compare our approach, AiR, with three existing unsupervised prompt-tuning methods: Few Pseudolabels (FPL) [21], Grow and Refine Iteratively Pseudolabels (GRIP) [21], and CPL [33].

Evaluation metrics. We evaluate each method by calculating test set accuracy and averaging results over three runs for consistency. In the case of Transductive Zero-Shot Learning (TRZSL), we report the harmonic mean of accuracies for seen and unseen classes to address potential performance imbalances between them.

Implementation details. In all experiments, spanning various datasets and learning strategies, we use ViT-B/32 as the vision backbone. We use Stable Diffusion v1-4 as our synthetic data generation model. For both visual and textual prompt tuning, the prefix size is set to 16. We employ SGD as the optimizer and train for 150 epochs, with an initial learning rate of 0.0001 during the first 5 warmup epochs, after which it is adjusted to l and decayed according to a cosine annealing schedule. For both textual and visual prompt tuning, l is set to 0.1. The number of iterations is set to 10. For FPL, the number of pseudo-labels per class is fixed at 16, following [21]. In GRIP and CPL, the number of pseudo-labels is treated as a tunable hyperparameter following [33].

4.2. Comparison with SOTA Methods

To validate the effectiveness of our proposed AiR, we employ CPL [33] as a baseline and conduct comparative experiments using three learning paradigms: SSL, UL, and TRZSL on five distinct datasets, benchmarking against contemporary SOTA methods. To demonstrate the generalizability of our approach, we evaluate AiR in both text prompt learning and visual prompt learning frameworks. As shown in Table 1 our approach consistently achieves SOTA results across various datasets and learning paradigms. In UL,

TP		RESISC45			FGVCAircraft			Flowers102			EuroSAT			DTD			Average		
Method	Year	SSL	UL	TRZSL	SSL	UL	TRZSL	SSL	UL	TRZSL	SSL	UL	TRZSL	SSL	UL	TRZSL	SSL	UL	TRZSL
CLIP[24]	2021	54.5	54.5	54.5	17.6	17.6	17.9	63.7	63.7	63.4	32.8	32.8	30.5	43.2	43.2	43.4	42.3	42.3	41.9
CoOp[36]	2022	58.5	-	63.4	14.9	-	21.7	76.8	-	63.2	59.5	-	49.6	37.1	-	46.3	49.3	-	48.8
FPL[21]	2023	68.1	63.0	72.1	20.0	16.6	17.5	75.9	65.6	80.9	62.0	48.9	53.7	37.1	44.9	46.3	52.6	47.8	54.0
GRIP[21]	2023	74.1	70.6	81.1	17.0	15.2	26.1	83.6	69.8	86.8	58.6	57.2	92.3	56.0	46.0	65.3	57.8	51.7	70.3
CPL[33]	2024	80.9	77.3	85.8	22.4	18.3	30.8	89.6	72.9	87.3	75.4	67.2	93.7	61.2	51.9	68.0	65.9	57.5	73.1
Ours	-	**82.5**	**79.9**	**87.5**	**24.2**	**19.6**	**33.2**	**91.2**	**74.3**	**92.3**	**76.4**	**71.4**	**93.9**	**62.2**	**55.7**	**69.9**	**67.3**	**60.1**	**75.4**
VP		RESISC45			FGVCAircraft			Flowers102			EuroSAT			DTD			Average		
CLIP[24]	2021	54.5	54.5	54.5	17.6	17.6	17.9	63.7	63.7	63.4	32.8	32.8	30.5	43.2	43.2	43.4	42.3	42.3	41.9
VPT[13]	2022	60.8	-	67.1	17.7	-	26.6	63.7	-	64.7	47.1	-	62.2	36.4	-	44.1	45.1	-	52.9
FPL[21]	2023	65.1	62.2	67.8	20.1	18.2	16.2	67.0	65.5	71.9	52.4	48.7	68.6	47.6	47.6	52.4	50.4	48.4	55.3
GRIP[21]	2023	71.2	68.4	82.1	19.4	17.5	26.4	67.9	63.0	77.1	63.4	63.6	96.9	54.5	50.5	62.7	55.2	52.6	68.4
CPL[33]	2024	78.4	72.9	86.6	20.5	18.2	30.2	73.5	67.2	80.1	72.0	68.9	98.3	58.7	53.4	65.3	60.6	56.1	72.1
Ours	-	**81.1**	**79.3**	**87.8**	**23.7**	**19.3**	**31.2**	**78.5**	**69.2**	**81.9**	**73.5**	**71.6**	**98.5**	**59.6**	**54.6**	**66.4**	**63.2**	**58.8**	**73.2**

Table 1. Comparison results of top-1 test accuracy(%) on five benchmarks. TP and VP represent applying text prompts and visual prompts as tuning strategies, respectively. The average accuracy on all datasets is shown on the right. The best results are in **bold**.

AiR enhances performance by an average of approximately 2.6% in text prompt learning and 2.7% in visual prompt learning. In both SSL and TRZSL, AiR achieves performance improvements ranging from 1% to 3%. These results indicate that AiR consistently boosts the classification performance of prompt learning models across different learning paradigms. This underscores the robustness and generalizability of our approach, demonstrating that augmenting discriminative richness effectively enhances unsupervised prompt learning performance in diverse scenarios. We also conducted experiments on the large-scale dataset and cross-dataset settings, please refer to the Appendix.

4.3. Ablation Studies

Effect of Synthetic Prediction and Synthetic Loss. We conduct ablation studies on three datasets using the UL and text prompt paradigm to demonstrate the individual contributions of each core component in our AiR method. In Table 2, $\hat{p}_c$ denotes the inclusion of discriminative results from synthetic samples in Eq. 7, $\mathcal{L}_s$ indicates whether these synthetic samples contribute to the loss calculation in Eq. 9, and $\mathcal{L}_r$ indicates whether real samples contribute to the loss calculation in Eq. 8. To isolate our approach's impact, we use GRIP [21] as the baseline and report model accuracy under each condition. When real samples loss function $\mathcal{L}_r$ is employed, using only $\hat{p}_c$ yields a 1.7%-5.4% improvement, indicating that enriching visual perception discriminative information has a substantial effect on model accuracy. Similarly, training with synthetic samples from the SD model together with real samples enhances performance by 1.5%-4.8%, showing that sample diversity and fidelity provide valuable information to the unsupervised model, helping it adapt prompts better suited to downstream tasks. However, given that our AiR model leverages synthetic samples generated by the SD model for the unsupervised prompt learning task, an important question arises: if only synthetic samples are used for training, is it feasible to fully address the unsupervised learning task? As the result shows, the model's performance declines significantly when no real samples are included in the training, for instance, a decrease of -10.1% in Flowers102 and -12.3% in RESISC45. These results indicate that relying solely on synthetic samples is insufficient for unsupervised learning. Although synthetic samples exhibit sufficient diversity and real-world fidelity, they still cannot fully replace the information provided by real samples. When $\hat{p}_c$, $\mathcal{L}_s$, and $\mathcal{L}_r$ are combined, the model achieves optimal performance, highlighting that the enhanced visual perception and sample diversity effects are complementary rather than redundant, yielding a compounded improvement. We also conducted experiments on the effect of different generation models, please refer to the Appendix for more details.

$\mathcal{L}_r$	$\hat{p}_c$	$\mathcal{L}_s$	RESICS45	Flowers102	DTD
✓			70.6	69.8	46.0
		✓	58.3	59.7	41.4
✓	✓		72.3	71.6	51.4
✓		✓	72.9	71.3	50.8
✓	✓	✓	**73.6**	**71.9**	**52.3**

Table 2. Accuracy scores (%) on RESISC45, Flowers102, and DTD are reported to verify the effects of different strategies. $\hat{p}_c$ represents the auxiliary classifier from synthetic samples, $\mathcal{L}_s$ and $\mathcal{L}_r$ indicate the training loss about the synthetic samples and real samples, respectively.

Effect of LoRA. To assess the necessity of LoRA for training SD models, we evaluate the accuracy of SD models with and without LoRA on three datasets (RESISC45, Flowers102, EuroSAT, and DTD) in the unsupervised paradigm, with CPL as the baseline. As illustrated in Table 3, synthetic samples generated with LoRA fine-tuning provide our AiR model with a 3%-8% accuracy gain compared to the case without LoRA in the same training environment. This result highlights that while SD can produce diverse and high-quality synthetic samples, alignment with the dataset's domain is crucial, making LoRA an indispensable component of our AiR model. We also conduct ablation experiments on cosine similarity for the sample selected, please refer to the Appendix for more details.

4.4. Exploratory Analysis

Sensitivity to the Number of Synthetic Samples. To investigate the impact of varying numbers of synthetic sam-

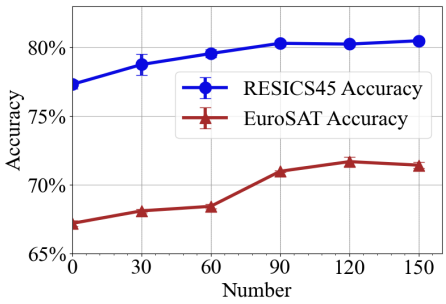

Figure 3. Comparison of top-1 test accuracy (%) across varying numbers of synthetic samples in unsupervised learning on the RESISC45 and EuroSAT datasets.

LoRA	RESISC45	Flowers102	EuroSAT	DTD
w/o	76.5	66.3	66.2	51.6
w	**79.9**$^{+3.4}$	**74.3**$^{+8.0}$	**71.4**$^{+5.2}$	**55.7**$^{+4.1}$

Table 3. Comparison results of top-1 test accuracy (%) on unsupervised learning with or without LoRA fine-tuning. Note that 'w/o' denotes the synthetic sample generated without LoRA.

ples on model performance, we conducted experiments assessing changes in model accuracy with 0–150 synthetic samples per class on the RESISC45 and EuroSAT datasets, respectively. As shown in Fig. 3, the trend line illustrates how model accuracy shifts with the increasing number of synthetic samples. We observe that model performance gradually improves as the number of synthetic samples rises, especially when increasing from 60 to 90 samples: EuroSAT accuracy improves by 3%, and RESISC45 accuracy by 1%. However, this improvement is not indefinite; beyond 90 synthetic images, performance gains slow, and after surpassing 120 samples, the trend stabilizes. This suggests that synthetic samples indeed enrich discriminative information for unsupervised prompt learning and in turn enable noticeable improvement of the model's performance, though there is a limit to this benefit. Based on these observations, we set the number of synthetic samples at 120 across all learning scenarios in our AiR framework. More experiments about the hyper-parameter λ in Eq. 7 and β in Eq. 10 are provided in the Appendix.

Pseudo Label Accuracy. To demonstrate the extent to which the auxiliary classifier, constructed with synthetic samples, improves the pseudo-labeling quality of the model's original text classifier. We use the original prompt "a photo of a" and compare the accuracy of pseudo-labeling for the top 50 maximum confidence samples. We employ synthetic samples generated by our AiR model as a reference across three different datasets. As shown in Fig. 4, leveraging synthetic sample-assisted classification within our AiR model results in at least a 3% improvement in the original text classification accuracy, with a 6% gain on RESISC45. This indicates that augmenting the discriminative

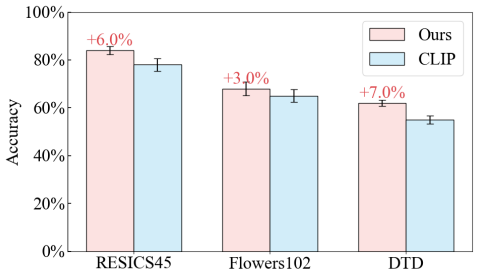

Figure 4. Comparison of top-1 test accuracy (%) of pseudo labels in unsupervised learning. 'Ours' denotes the use of our approach AiR. while 'CLIP' indicates the use of CLIP's text encoder.

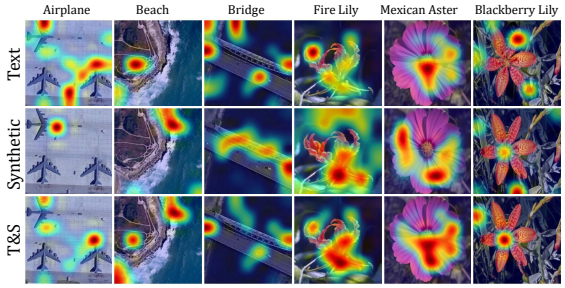

Figure 5. Visualization of the activation regions for the text classifier and synthetic-sample-based auxiliary classifier using CAM. 'T & S' represents the activation after merging the outputs of both classifiers.

richness through enhanced visual perceptual information significantly improves pseudo-labeling accuracy, which in turn boosts the performance of unsupervised prompt learning models. More experiments about pseudo-label accuracy are provided in the Appendix.

4.5. Qualitative Analysis

CAM of Different Discrimination Outcomes. To show the effectiveness of synthetic image embeddings intuitively, we use CLIP [24] as a baseline and perform CAM (Class Activation Mapping) [34] visualization on two datasets. As shown in Fig. 5, we observe that text embeddings sometimes focus on class-irrelevant regions, such as airports in the *Airplane* class, seawater in *Bridge*, and grass in *Blackberry Lily*. Additionally, text embeddings may fail to capture all relevant areas, such as part of a *Beach* or a flower petal in *Mexican Aster*. When using synthetic samples for discrimination, similar issues arise, but the regions of attention differ. For instance, while text embeddings overly focus on airports in *Airplane*, synthetic samples pay less attention to class-relevant regions of the airplane. Similarly, text embeddings focus on grass in *Blackberry Lily*, while synthetic samples shift focus to the location of stamens. This suggests a fundamental difference in the type of information emphasized by different discriminative modalities for the same sample. Moreover, when combining the

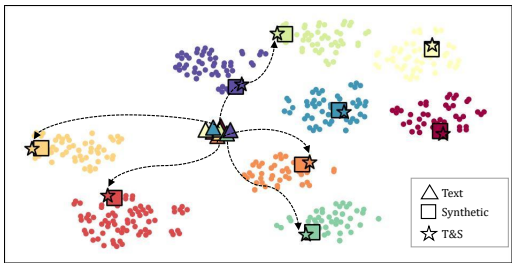

Figure 6. Visualization of the spatial distribution of synthetic image, text, and sample embeddings with t-SNE, respectively. Circular dots in different colors represent distinct image classes in Flowers102, triangles indicate text embeddings for each class, squares represent synthetic image embeddings, and pentagrams denote the embeddings after fusing text and synthetic image embeddings.

discriminative information from both text and synthetic images, we find that the fused results provide a more complete focus on class-relevant regions while minimizing the influence of class-irrelevant areas. For example, the fused attention highlights larger regions of the beach in *Beach* and more complete flowers in *Fire Lily* and *Mexican Aster*. This result confirms why our AiR method works because when the visual perceptual information embedded in the synthetic samples is used to enhance the richness of the discrimination, it helps the model to focus more on class-relevant regions and suppress activation in class-irrelevant areas, thus resulting in an improvement in model performance.

t-SNE of Different Discrimination Outcomes. In the previous section, we demonstrated the differences in image activation regions across various discriminative methods using CAM. In this section, we examine the spatial distribution of synthetic image embeddings, text embeddings, and sample embeddings through t-SNE visualization. As shown in Fig. 6, circular dots in different colors represent distinct image classes in Flowers102, triangles indicate text embeddings for each class, squares represent synthetic image embeddings, and pentagrams denote the embeddings after fusing text and synthetic image embeddings as outlined in Sec. 3.5. From the figure, it is apparent that due to modality differences, text embeddings (triangles) collapse into a compact region that is only loosely aligned with the corresponding class's test samples, rather than overlapping with them. Conversely, the synthetic sample embeddings (squares) align directly with data clusters for each class, appearing more suitable as classifiers compared to the text embeddings. However, the observed tight clustering of text embeddings in two-dimensional t-SNE space is primarily due to modality discrepancies, which may not reflect their true spatial distribution in higher-dimensional space. Thus, we cannot conclude outright that synthetic image embeddings are categorically more suitable for classification. Nevertheless, by fusing the embeddings (pentagrams), we see a clear tendency for text embeddings to move toward their corresponding test sample classes. This shift suggests that augmenting discriminative information effectively calibrates the embeddings, aligning them more closely with the correct test samples and thus improving the model's classification accuracy across different classes. More experiments on t-SNE are provided in the appendix.

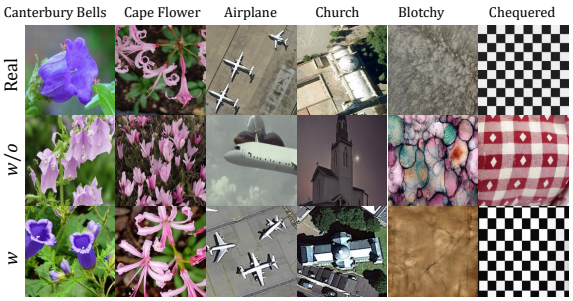

Figure 7. Visualization of real images and synthetic images generated by the SD model, with and without fine-tuning via LoRA.

Fidelity of generated samples w and w/o LoRA. We provide examples of images generated by the SD model with and without LoRA to visually demonstrate the importance of LoRA in our AiR model. As shown in Fig. 7, images generated without LoRA bear a general resemblance to the real images, yet differ significantly upon closer inspection. For instance, the petal shapes in *Canterbury Bells*, the viewpoint in *Airplane*, and the texture details in *Blotchy* reveal substantial mismatches; without LoRA, the SD model can only approximate the overall class, lacking finer, more precise details. In contrast, with LoRA applied, the generated images exhibit more accuracy in form and detail: the petal shapes in *Canterbury Bells* closely match the real image, the airplane's viewpoint in *Airplane* shifts to a top view, and the texture in *Blotchy* appears smoother and more aligned. These improvements demonstrate that without LoRA, synthetic samples lack sufficient detailed expression to real samples and cannot effectively serve as classification embeddings to augment discriminative information. This underscores the necessity of LoRA in our AiR model.

5. Conclusion

In this paper, we introduce AiR, a novel auxiliary classification approach designed to leverage synthetic samples in building an auxiliary classifier. In this way, we can capture richer visual variations while preserving core semantics, effectively bridging text-image-pair classification to image-image-pair classification. Extensive experiments across five benchmark datasets and three learning paradigms validate the effectiveness of our approach. We hope this work inspires new approaches to reducing dependence on labeled data when adapting pre-trained vision-language models like CLIP to novel tasks.

Acknowledgements. This work of Dandan Guo was supported by the National Natural Science Foundation of China (NSFC) under Grant 62306125. This work of Yi Chang was supported in part by the National Key R&D Program of China under Grant 2023YFF0905400, in part by the National Natural Science Foundation of China under Grant U2341229. This work of Fan Tang was supported in part by the Beijing Science and Technology Plan Project under no. Z231100005923033, and in part by Beijing Natural Science Foundation under no. L221013.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. 3

[2] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022. 3

[3] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[4] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 5

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5

[6] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023. 3, 4

[7] 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. 1

[8] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2, 4

[11] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 1, 2, 3

[12] 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 *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[13] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 1, 5, 6

[14] Yunbing Jia, Xiaoyu Kong, Fan Tang, Yixing Gao, Weiming Dong, and Yi Yang. Revealing the two sides of data augmentation: an asymmetric distillation-based win-win solution for open-set recognition. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 911–919, 2024. 1

[15] Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. Datadream: Few-shot guided dataset generation. *arXiv preprint arXiv:2407.10910*, 2024. 3

[16] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 3

[17] Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16155–16165, 2023. 1

[18] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022. 3

[19] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 3

[20] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[21] Cristina Menghini, Andrew Delworth, and Stephen Bach. Enhancing CLIP with CLIP: Exploring pseudolabeling for limited-label prompt tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 3, 5, 6

[22] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 2, 5

[23] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[24] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6, 7

[25] Hairui Ren, Fan Tang, Xingjia Pan, Juan Cao, Weiming Dong, Zhiwen Lin, Ke Yan, and Changsheng Xu. A^2Pt: Anti-associative prompt tuning for open set visual recognition. *IEEE Transactions on Multimedia*, 26:8419–8431, 2024. 1

[26] Hairui Ren, Fan Tang, Huangjie Zheng, He Zhao, Dandan Guo, and Yi Chang. Modality-consistent prompt tuning with optimal transport. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3):2499–2512, 2025. 1

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[28] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 1

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[30] Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *International Conference on Machine Learning*, pages 33816–33832. PMLR, 2023. 1

[31] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022. 3

[32] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1

[33] Jiahan Zhang, Qi Wei, Feng Liu, and Lei Feng. Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data. In *Forty-First International Conference on Machine Learning*, 2024. 1, 2, 3, 5, 6

[34] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 7

[35] 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 (CVPR)*, pages 16816–16825, 2022. 1, 2

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, pages 1–12, 2022. 1, 2, 5, 6