Supplementary Material for "Beyond Words: Augmenting Discriminative Richness via Diffusions in Unsupervised Prompt Learning"

A. Overall Algorithm

We present the comprehensive algorithmic flow of the AiR method in Algorithm 1. To balance data diversity and domain consistency in augmented images, we first fine-tune the SD model using LoRA. Subsequently, synthetic images $\hat{X}$ are generated via the fine-tuned Stable Diffusion Model (SD), and the samples most aligned with the semantic information are selected based on cosine similarity. The chosen synthetic images $\hat{X}^*$ serve as auxiliary classifiers, working alongside text classifiers to assign pseudo-labels $\tilde{y}^*$ to the unlabeled samples. Finally, the pseudo-labeled data $(x, \tilde{y}^*)$ and synthetic data $(\hat{x}, y)$ are used as training samples to compute the loss and iteratively optimize the prompt.

Algorithm 1: Workflow about our AiR method for optimizing prompt t

Get pseudo label data $\{x, \tilde{y}\}$ in Eq. (2)
Finetune SD with LoRA in Eq. (4)
Generate synthetic images $\hat{X}$ in Eq. (5)
Select high confidence samples $\hat{X}^*$
for $iter = 1, 2, 3, ...$ **do**
  Text-image prediction $p_c = \frac{\exp(\text{sim}(f, g_c)/\tau)}{\sum_{c'=1}^{C} \exp(\text{sim}(f, g_{c'})/\tau)}$
  Image-image prediction $\hat{p}_c = \frac{\exp(\text{sim}(f, \hat{x}_c^*)/\tau)}{\sum_{c'=1}^{C} \exp(\text{sim}(f, \hat{x}_{c'}^*)/\tau)}$
  Combine both types of predictions $p_c^* = p_c + \lambda * \hat{p}_c$
  Pseudo labels $\tilde{y}^* = \underset{c}{\text{argmax}} p_c^*$
  Real data loss $\min_t \mathcal{L}_r = \min_t \mathcal{L}_{ce}(x, \tilde{y}^*)$
  Synthetic data loss $\min_t \mathcal{L}_s = \min_t \mathcal{L}_{ce}(\hat{x}, y)$
  Total loss $\min_t \mathcal{L} = \min_t (\mathcal{L}_r + \beta * \mathcal{L}_s)$
  Take gradient descent step on $\nabla \mathcal{L}$
  Update $t^{iter+1} \leftarrow t^{iter} - \eta \nabla \mathcal{L}$
end

Method	RESISC45	EuroSAT	DTD
Kandinsky2.2	77.6	69.5	54.3
Dreambooth	78.3	70.8	54.8
Stable Diffusion	**79.9**	**71.4**	**55.7**

Table S1. Comparison of top-1 test accuracy (%) in unsupervised learning with different generative models: Kandinsky2.2, Dreambooth, and Stable Diffusion. The best results are in **bold**.

B. Task Introduction

We outline the experimental settings for three distinct learning paradigms following [4, 9]:

Semi-Supervised Learning (SSL): In this paradigm, access to labeled data is restricted. To evaluate the influence of pseudo-labels, we consider scenarios with minimal labeled data and abundant unlabeled data, using two labeled samples per class.

Unsupervised Learning (UL): Here, only unlabeled data is available. Pseudo-labels are initially derived entirely from the zero-shot predictions of CLIP, eliminating the need for any manual annotation.

Transductive Zero-Shot Learning (TRZSL): In this setting, labeled data is provided for a subset of target classes (seen classes) in the downstream dataset. We set the seen-to-unseen class ratio to 62:38, with pseudo-labels generated exclusively for unseen classes. Additionally, for TRZSL, we report the harmonic mean of the accuracies for seen and unseen classes to account for performance balance.

C. Finetune with LoRA

We fine-tune the Stable Diffusion-v1-4 (SD) model using LoRA [3], following the approach in [8]. For each dataset, we select the top 5 pseudo-labeled samples with the highest confidence to train the SD model. To ensure alignment between the SD model and the semantic space of CLIP's text encoder, only the U-Net component of SD is fine-tuned. The model is trained for 15,000 steps with a batch size of 1, using a learning rate of 1e-5.

Method	Flowers102	DTD	EuroSAT
Euclidean Distance	54.8	51.5	73.0
OT distance	68.3	62.3	76.6
Cosine Distance	**68.3**	**62.3**	**76.6**

Table S2. Comparison of top-1 test accuracy (%) in unsupervised learning with different similarity metrics: Euclidean Distance, OT distance, and Cosine Distance. The best results are in **bold**.

D. Effect of Generative Models

To investigate the performance variations of our AiR method across different generative models, we fine-tune three models: Kandinsky 2.2, DreamBooth [6], and Stable Diffusion using LoRA and apply the same training strategy for the AiR model, with CPL as the baseline. As shown in Table S1, we conduct experiments on three datasets, and the results indicate that the performance of AiR across these generative models varies by less than 2%. This demonstrates that our method is not heavily dependent on the choice of the generative model, as long as the model can be fine-tuned to ensure both the fidelity of the generated images and sample diversity. However, DreamBooth requires individual fine-tuning for each category, leading to higher computational costs, while Kandinsky 2.2 shows slightly lower performance compared to Stable Diffusion. Consequently, we select Stable Diffusion as the generative model for our AiR framework.

E. Effect of Cosine Selected Strategy

To determine the most effective filtering strategy for selecting representative synthetic samples, we evaluate different similarity metrics between synthetic samples and textual features across three datasets. We report the pseudo-labeling accuracies of the Top-50 confidence samples for the CLIP model after incorporating synthetic samples as auxiliary classifiers. As shown in Table S2, we test the Euclidean distance, OT distance [7], and cosine similarity. The Euclidean distance results in a significant drop in performance, indicating its unsuitability for selecting synthetic samples. Both OT distance and cosine similarity demonstrate comparable performance, with a slight degradation observed for OT distance. Considering that OT distance requires higher computational resources compared to cosine similarity, we ultimately select cosine similarity as the filtering strategy to identify the most representative synthetic samples.

F. Hyper-parameters

To investigate the influence of hyperparameters λ and β discussed in Sec. 3.5, we evaluate the accuracy of pseudo-labels by varying their values on the EuroSAT [2] and RESISC45 [1] datasets. As illustrated in Fig. S1 and Fig. S2, the model's accuracy fluctuates by approximately 3% when λ changes from 1/8 to 1. This suggests that effectively balancing the results of the auxiliary classifier (constructed using synthetic samples) and the text classifier is crucial. Overemphasizing either side diminishes the quality of the pseudo-labels. We observe optimal performance when λ is set to 1/6 or 1/4, achieving 78.6% accuracy on the EuroSAT dataset and 83.8% on the RESISC45 dataset. Consequently, for our final experiments, we select λ as 1/6. Similarly, when adjusting β between 1/4 and 2, the model's performance remains relatively stable. This indicates that optimizing the network with synthetic samples as a loss function does not require extensive fine-tuning of β. Ultimately, we choose 1 as the value of β for its higher performance consistency.

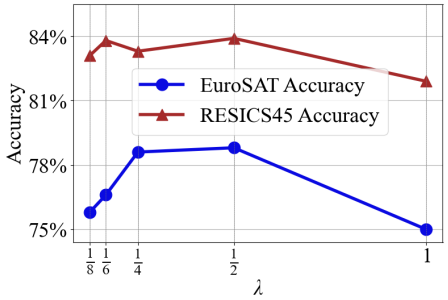

Figure S1. Comparison of top-1 test accuracy (%) of pseudo labels with different hyper-parameters λ.

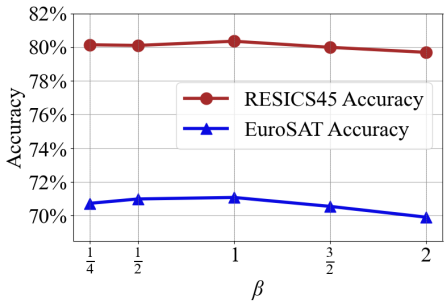

Figure S2. Comparison of top-1 test accuracy (%) of pseudo labels with different hyper-parameters β.

G. Pseudo Label Accuracy

To further illustrate the impact of our method on enhancing pseudo-labeling quality, we compare the pseudo-labeling accuracy of CPL (used as a baseline) with that of our AiR method on the Flowers102 [5] dataset over successive training iterations. As depicted in Fig. S3, the pseudo-

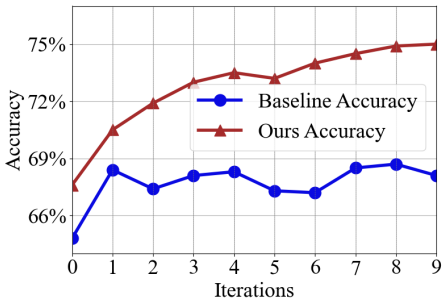

Figure S3. Comparison of top-1 test accuracy (%) of pseudo labels during different training iterations.

labeling accuracy of our method consistently surpasses that of the baseline by 2%-7% as training progresses. This highlights the sustained improvement in pseudo-labeling quality achieved by our approach. The stable accuracy margin of over 2% further validates that our method generates high-quality pseudo-labeled samples, enabling the training of a more robust unsupervised prompt learning model.

H. t-SNE of Different Discrimination Outcomes.

In this section, we analyze the spatial distribution of synthetic image embeddings, text embeddings, and sample embeddings using t-SNE visualizations on the DTD and RESISC45 datasets. As illustrated in Fig. S4 and Fig. S5, colored circular dots represent different image classes, triangles indicate the text embeddings for each class, squares denote the synthetic image embeddings, and pentagrams represent the fused embeddings of text and synthetic images, as described in Sec. 3.5. The findings are consistent with the results discussed in the main paper. The fused embeddings (pentagrams) reveal a clear tendency for text embeddings to shift toward their corresponding test sample classes. This alignment indicates that augmenting discriminative information effectively calibrates the embeddings, bringing them closer to the correct test samples and enhancing the model's classification accuracy across various classes.

I. Comparison on Large Scale Dataset.

We conduct experiments on ImageNet, which is challenging and not considered by other works. Due to memeory constraints, we select 100 samples per class. As shown in Table S3, ours outperforms CPL by nearly 2% accuracy. These results indicate that our AiR consistently boosts the classification performance of prompt learning models even on large scale dataset, demonstrate the robustness of our method.

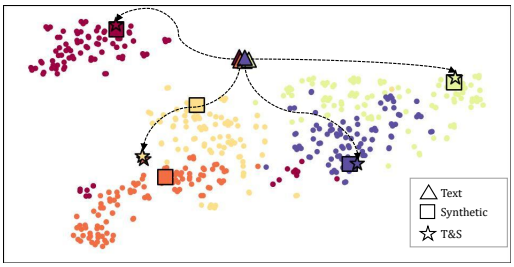

Figure S4. Visualization of the spatial distribution of synthetic image, text, and sample embeddings with t-SNE in DTD dataset.

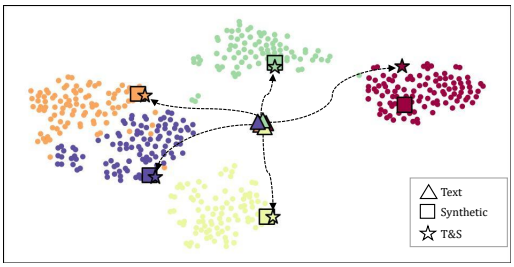

Figure S5. Visualization of the spatial distribution of synthetic image, text, and sample embeddings with t-SNE in RESISC45 dataset.

Method	ImageNet		
	SSL	UL	TRZSL
CPL	61.6	62.8	65.3
Ours	**63.9**	**64.9**	**67.2**

Table S3. Comparison results of top-1 test accuracy (%) on ImageNet dataset. The best results are in **bold**.

Method	Source	Target				
	RESISC45	Flowers102	FGVCaircraft	DTD	EuroSAT	Average
CPL	77.3	36.2	2.6	16.9	35.3	22.7
Ours	**79.9**	**49.6**	**6.9**	**24.3**	**44.5**	**31.3**
	Flowers102	RESISC45	FGVCaircraft	DTD	EuroSAT	Average
CPL	72.9	20.4	6.3	20.9	34.7	20.5
Ours	**74.3**	**28.4**	**6.5**	**29.3**	**39.9**	**26.0**
	DTD	RESISC45	FGVCaircraft	Flowers102	EuroSAT	Average
CPL	51.9	40.8	4.4	45.7	27.5	29.6
Ours	**55.7**	**44.3**	**5.2**	**47.9**	**32.5**	**32.5**

Table S4. Results in cross-dataset scenarios.

J. Comparison in Cross-dataset Scenarios.

To further demonstrate the generalisability of our approach, we perform comparative experiments in cross-dataset scenarios. We train models on RESISC45/Flowers102/DTD and test them on other datasets. As shown in Tab. S4, our Air still surpasses the SOTA-CPL by **3-8%** accuracy, proving its generalizability.

References

[1] 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. 2

[2] 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. 2

[3] 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. 1

[4] 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

[5] 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

[6] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[7] Cédric Villani et al. *Optimal transport: old and new*. Springer, 2009. 2

[8] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 1

[9] 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