A Appendix

We include here extra information that supports the results presented in the main body of the paper.

Reproducibility We have provided the code to run the experiments as supplementary material for the submission. However, we plan to release it as an open repository upon acceptance.

A.1 Trainable Prompts

Text Prompt Tuning The primary objective of text prompt tuning is to improve the alignment between the class token and the image features extracted by the image encoder. This is achieved by adding learnable vectors, i.e., *prefix*, before the `CLASS` token to create a contextualized representation. Specifically, the sequence

$$\boldsymbol{t} = [\mathrm{V}]_1[\mathrm{V}]_2 \dots [\mathrm{V}]_M[\mathrm{CLASS}]$$

is fed into the textual encoder, where each vector $[\mathrm{V}]_m$ $(m \in 1, \dots, M)$ has the same dimension as word embeddings, and M is a hyperparameter that determines the length of the prefix.

Context Optimization (CoOp) [48] was the first work to explore continuous prompts for VLMs. Follow-up works have experimented with different training strategies to enhance the generalizability of the learned prompts while preserving the core concept of continuous vector tuning [34, 12, 27, 46, 13, 37].

Tuning the text prefix vector changes the resulting n linear weight vectors $w_i = \psi(p_i)$, while leaving the image features unchanged. Therefore, text prompt tuning may be most beneficial when image features are well-separated by class but may not be aligned with the corresponding textual prompt. Conversely, text prompt tuning may not be as effective when the image features are poorly separated, as in specialized or novel domains where CLIP may lack sufficient training data.

Visual Prompt Tuning Instead of tuning the text prompts, one can also tune the inputs of the vision encoder. In this case, a learnable visual prefix is prepended to the image tokens as input to the image transformer as follows:

$$\hat{\boldsymbol{I}} = [\mathrm{p}]_1 \dots [\mathrm{p}]_K[\mathrm{I}]_1 \dots [\mathrm{I}]_P$$

where p represents a sequence of K learnable prefix vectors, and $[\mathrm{I}]_1 \dots [\mathrm{I}]_P$ are the image tokens from the corresponding P patches of the input images. The new sequence $\hat{\boldsymbol{I}}$ is the input to the image encoder ϕ.

Visual Prompt Tuning (VPT) was introduced in the context of efficiently adapting pre-trained vision transformers to downstream tasks [18]. However, the approach has since been applied in the context of VLM [34].

Whereas text prompt tuning does not alter the image features, visual prompt tuning does. By rearranging the image features within the projection space, VPT has the potential to improve CLIP when the image features are not well separated by class, such as in specialized domains.

Multimodal Prompt Tuning The previous approaches are unimodal, as they either involve modifying the text or visual input, but never both. This choice may be suboptimal as it does not allow the flexibility to dynamically adjust both representations on a downstream task. Recently, multimodal prompt tuning has been introduced [44, 19]. We focus on Unified Prompt Tuning (UPT) [44] which essentially learns a tiny neural network to jointly optimize prompts across different modalities. UPT learns a set of prompts $\boldsymbol{U} = [\boldsymbol{U}_T, \boldsymbol{U}_V] \in \mathbb{R}^{d \times n}$ with length n, where $\boldsymbol{U}_T \in \mathbb{R}^{d \times n_T}, \boldsymbol{U}_V \in \mathbb{R}^{d \times n_V}$. $\boldsymbol{U}$ is transformed as follows:

$$\boldsymbol{U}' = \mathrm{SA}(\boldsymbol{U}) + \mathrm{LN}(\boldsymbol{U})$$
$$\hat{\boldsymbol{U}} = \mathrm{FFN}\left(\mathrm{LN}\left(\boldsymbol{U}'\right)\right) + \mathrm{LN}\left(\boldsymbol{U}'\right)$$

where SA is the self-attention operator, LN is the layer normalization operator, and FFN is a feed forward network. After transformation, we obtain $\hat{\boldsymbol{U}} = \left[\hat{\boldsymbol{U}_T}, \hat{\boldsymbol{U}_V}\right] \in \mathbb{R}^{d \times n}$, such that $\hat{\boldsymbol{U}_T}$ is to be used as a text prompt, and $\hat{\boldsymbol{U}_V}$ is to be used as a visual prompt.

	Num. classes ($\lvert\mathcal{Y}\rvert$)	Num. seen classes ($\lvert S\rvert$)	Num. unseen classes ($\lvert U\rvert$)	Size training data	Avg. labeled data per class	Size test
Flowers102	102	63	39	2040	16	6149
RESICS45	45	27	18	6300	110	25200
FGVC-Aircraft	100	62	38	6667	53	3333
MNIST	10	6	4	60000	4696	10000
EuroSAT	10	6	4	27000	2200	5000
DTD	47	29	18	3760	64	1880

Table 4: For each dataset we report the number of classes, the number of seen and unseen classes in the TRZSL setting, the size of training data (including both labeled and unlabeled data), the average number of labeled examples per class, and the size of the test set which is the same across learning paradigms. We recall that we use the datasets gathered by the recent ELEVATER [23] benchmark for vision-language models.

The author of UPL argue that self-attention allows for beneficial interaction between the two separate modalities, which leads to both separable visual features, and text classifiers that are well-aligned with the corresponding visual features [44].

Prompts initialization We initialize textual and visual prompts from a normal distribution of mean 0 and variance 0.02. We note that we learn shallow visual prompts by modifying only the input to the image encoder. Multimodal prompts are initialized from a uniform distribution. We found that the latter was not working properly for textual and visual prompts.

Additional training settings For training, the batch size is 64.

A.2 Datasets details

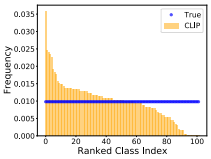
We use six datasets from specialized or fine-grained domains. Here we provide a description of each of them. In Table 4, we report the details about the number of classes and data available for each dataset. For each dataset, we also show CLIP's prediction distribution over classes Figure 5.

Flowers102 [29] It is a dataset collecting images for 102 flower categories commonly occurring in the United Kingdom. For each class we have between 40 and 258 images. Figure 5a shows that CLIP's predictions are skewed toward certain classes, which are predicted more often than what we would expect according to the real class distribution on the test set.

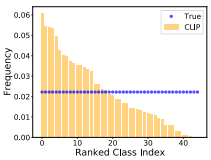
RESICS45 [9] This is a publicly available benchmark for Remote Sensing Image Scene Classification. It collects 45 kind of scenes. Figure 5b shows that CLIP predicts more often a subset of classes.

FGVC-Aircraft [26] It describes the fine-grained task of categorizing aircraft. We consider the task of classifying aircrafts into 100 variants. Also for this task, CLIP assigns images to a reduced set of classes (Figure 5c).

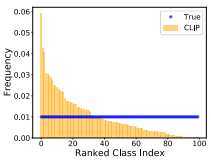
MNIST [11] MNIST is a database of handwritten digits. The digits are size-normalized and centered in a fixed-size image. We observe that CLIP never predicts 6 out of 10 classes (Figure 5c).



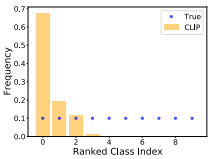
(a) Flowers102



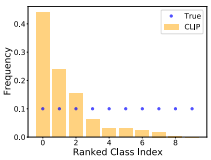
(b) RESICS45



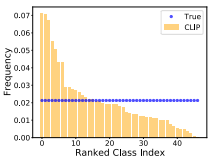
(c) FGVC-Aircraft



(d) MNIST



(e) EuroSAT



(f) DTD

Figure 5: For each dataset we show the distribution of CLIP's predictions over classes on the test set. The blue dots represent the true class distribution.

Multimodal prompts									
	Flowers102			RESICS45			FGVCAircraft		
Method	SSL	UL	TRZSL	SSL	UL	TRZSL	SSL	UL	TRZSL
CLIP	$63.67_{0.00}$		$63.40_{0.00}$	$54.48_{0.00}$		$54.46_{0.00}$	**17.58**$_{0.00}$		**17.86**$_{0.00}$
UPT	$68.03_{1.29}$	-	$61.05_{0.04}$	$62.84_{1.05}$	-	$58.79_{0.04}$	$11.13_{4.98}$	-	$15.89_{0.07}$
GRIP	**74.56**$_{2.02}$	$64.82_{1.63}$	**82.01**$_{0.01}$	**73.68**$_{0.91}$	$69.37_{0.61}$	**82.17**$_{0.00}$	$17.36_{0.43}$	$14.73_{0.08}$	**17.85**$_{10.30}$
Δ CLIP	↑ 10.89	↑ 1.15	↑ 18.61	↑ 19.2	↑ 14.89	↑ 27.71	↓ 0.22	↓ 2.85	↓ 0.01
Δ UPT	↑ 6.53	-	↑ 20.96	↑ 10.84	-	↑ 22.38	↑ 6.23	-	↑ 1.96
	MNIST			EuroSAT			DTD		
CLIP	$25.10_{0.00}$		$20.77_{0.00}$	$32.88_{0.00}$		$30.54_{0.00}$	$43.24_{0.00}$		$43.45_{0.00}$
UPT	**64.44**$_{3.66}$	-	$63.59_{0.11}$	**68.85**$_{9.92}$	-	$60.43_{0.04}$	$43.71_{2.18}$	-	$36.91_{0.04}$
GRIP	**65.94**$_{2.23}$	**68.18**$_{run}$	**73.75**$_{2.93}$	**60.38**$_{4.77}$	**61.52**$_{3.04}$	**95.52**$_{0.40}$	**54.07**$_{2.25}$	**47.37**$_{0.7}$	**63.42**$_{0.00}$
Δ CLIP	↑ 40.84	↑ 43.08	↑ 52.98	↑ 27.5	↑ 28.64	↑ 64.98	↑ 10.83	↑ 4.13	↑ 19.97
Δ UPT	↑ 2.35	-	↑ 10.16	↓ 8.47	-	↑ 35.09	↑ 10.36	-	↑ 26.51

Table 5: For each learning paradigm, we compare the accuracy of GRIP with CLIP zero-shot (ViT-B/32), and UPL. Results are for SSL, UL, and TRZSL on FRAMED. We average the accuracy on 5 seeds and report the standard deviation. Δ `METHOD` is the difference between the accuracy of GRIP and `METHOD`. We note that for UL we can not apply UPL since no labeled data is available.

EuroSAT [14] EuroSAT represents the task of categorizing satellite images of scenes. It consists of 10 classes. In Figure 5e, we show CLIP's predictions distribution over the classes.

DTD [10] DTD stands for Describable Textures Dataset. It is an evolving collection of textural images in the wild, and it is annotated relying on human-centric attributes, inspired by the perceptual properties of textures. The zero-shot CLIP predictions show the model's bias toward certain classes (Figure 5f).

A.3 Experiments

In this section, we report tables and plots that complement the results presented in Section 4.

The effect of GRIP on multimodal prompts Table 5 shows the improvements of GRIP on CLIP and Unified Prompt Tuning (UPL) [44]. Similar to the results in Table 1, GRIP consistently improves CLIP with respect to the baselines. The improvements on CLIP are by 18.2 in semi-supervised learning, 14.8 in unsupervised learning, and 30.7 in transductive zero-shot learning. While GRIP outperforms UPL by 4.7 in semi-supervised learning, and 19.5 in transductive zero-shot learning.

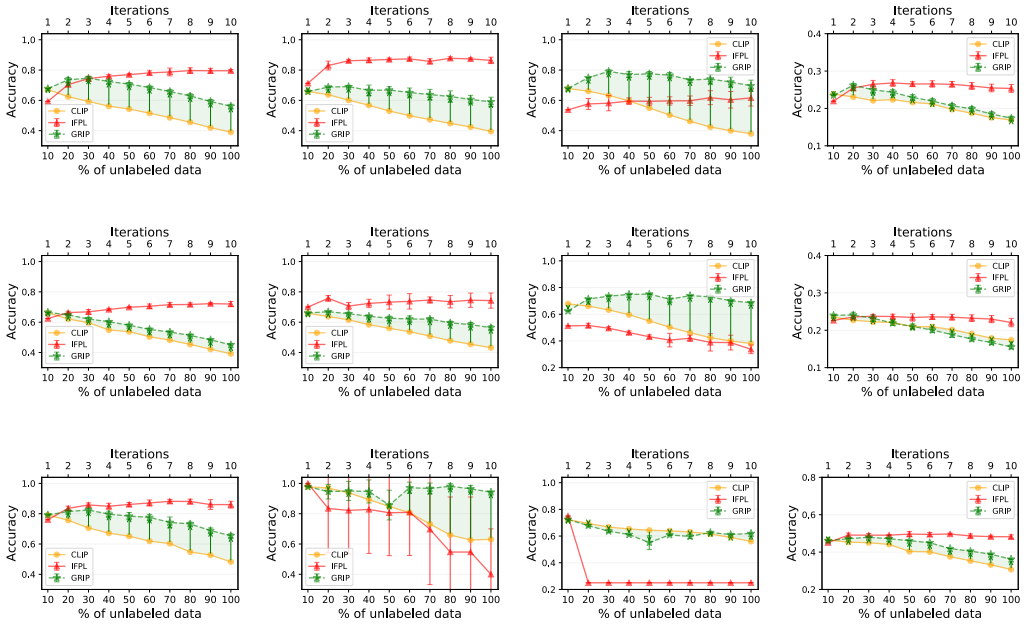
Comparison across iterative strategies In Table 6, we report a comparison between FPL and the iterative strategies (IFPL and GRIP) on MNIST, EuroSAT, and FGVC-Aircraft. Results on the other tasks can be found in the main body of the paper Section 4.1. While GRIP largely and consistently outperforms FPL by on average 16.7 points in accuracy, IFPL is not robust and it leads to performances that are inferior to FPL by on average 4.4 points in accuracy.

The evolving accuracy of dynamic pseudolabels Figure 6 represents the evolution of pseudolabels accuracy during training for all datasets, but Flowers102 and RESICS45 presented in Figure 3. We observe that the accuracy of the pseudolabels characterizes the overall performance of the models reported in Table 6. For instance, IFPL for EuroSAT in the TRZSL setting is highly variable, explaining the low average accuracy of the model on the test set (Table 6). Similarily, for MNIST in the TRZSL we observe that after the first iteration, the pseudolabels get very noisy.

GRIP performance on transductive zero-shot learning We show how the effectiveness of GRIP is consistent over the three random splits of seen and unseen classes which we randomly generated. The splits are reported in Table 9. Table 8 gathers the accuracy of seen and unseen classes, along with the harmonic mean for all three splits using textual prompts. Beyond the consistent improvement induced by GRIP training strategy, we observe that the accuracy of GRIP on the seen classes is often lower than the accuracy of CoOp on the same set of classes. We speculate this can result from two factors: (1) we learn to distinguish between seen and unseen losing knowledge specialized on the

Textual prompts									
	MNIST			EuroSAT			FGVCAircraft		
Method	SSL	UL	TRZSL	SSL	UL	TRZSL	SSL	UL	TRZSL
FPL	$66.06_{1.10}$	$40.03_{2.63}$	$9.73_{19.45}$	$\mathbf{62.05}_{1.64}$	$48.96_{1.49}$	$53.70_{26.87}$	$\mathbf{20.02}_{0.77}$	$\mathbf{16.62}_{0.67}$	$17.55_{0.37}$
IFPL	$59.14_{3.43}$	$28.94_{2.05}$	$0.00_{0.00}$	$\mathbf{61.28}_{1.59}$	$\mathbf{56.46}_{3.26}$	$14.36_{28.71}$	$18.00_{0.35}$	$13.80_{0.67}$	$21.72_{0.77}$
GRIP	$\mathbf{71.78}_{3.59}$	$\mathbf{67.88}_{2.76}$	$\mathbf{74.06}_{0.29}$	$58.66_{2.64}$	$\mathbf{57.21}_{1.77}$	$\mathbf{92.33}_{0.69}$	$16.98_{0.82}$	$15.22_{0.71}$	$\mathbf{26.08}_{0.25}$
Δ IFPL	↓ 6.92	↓ 11.09	↓ 9.73	↓ 0.77	↑ 7.50	↓ 39.34	↓ 2.02	↓ 2.82	↑ 4.17
Δ GRIP	↑ 5.72	↑ 27.85	↑ 64.33	↓ 3.39	↑ 8.25	↑ 38.63	↓ 3.04	↓ 1.40	↑ 8.53
Visual prompts									
FPL	$42.84_{16.80}$	$39.62_{6.53}$	$31.82_{17.53}$	$52.47_{2.53}$	$48.79_{3.69}$	$68.68_{14.74}$	$\mathbf{20.14}_{0.26}$	$\mathbf{18.28}_{0.33}$	$16.28_{0.45}$
IFPL	$52.91_{8.99}$	$37.17_{6.27}$	$38.38_{4.21}$	$\mathbf{57.85}_{6.52}$	$32.52_{10.00}$	$48.13_{11.13}$	$18.77_{0.48}$	$16.36_{0.37}$	$19.29_{0.36}$
GRIP	$\mathbf{69.66}_{5.51}$	$\mathbf{68.04}_{1.11}$	$\mathbf{69.54}_{1.31}$	$\mathbf{63.48}_{3.09}$	$\mathbf{63.68}_{3.42}$	$\mathbf{96.97}_{0.77}$	$\mathbf{19.43}_{0.50}$	$\mathbf{17.51}_{0.61}$	$\mathbf{26.42}_{0.30}$
Δ IFPL	↑ 10.07	↓ 2.45	↑ 6.56	↑ 5.38	↓ 16.27	↓ 20.55	↓ 1.37	↓ 1.92	↑ 3.01
Δ GRIP	↑ 26.82	↑ 28.42	↑ 37.72	↑ 11.01	↑ 14.89	↑ 28.29	↓ 0.71	↓ 0.77	↑ 10.14

Table 6: For each learning paradigm, we compare FPL, IFPL, and GRIP on MNIST, EuroSAT, and FGVCAircraft. We average across 5 runs and report the standard deviation. Δ METHOD is the difference between the accuracy of FPL and METHOD.



(a) DTD (b) EuroSAT (c) MNIST (d) FGVC-Aircraft

Figure 6: We plot the evolution of dynamic-pseudolabels accuracy during training. The rows refer to SSL, UL, and TRZSL, in order. IFPL refers to the top x-axis, while CLIP and GRIP to the bottom.

seen classes, and (2) the parameter λ that upweights the error on the pseudolabeled data is too large (Section 3.3) and further training might be needed.

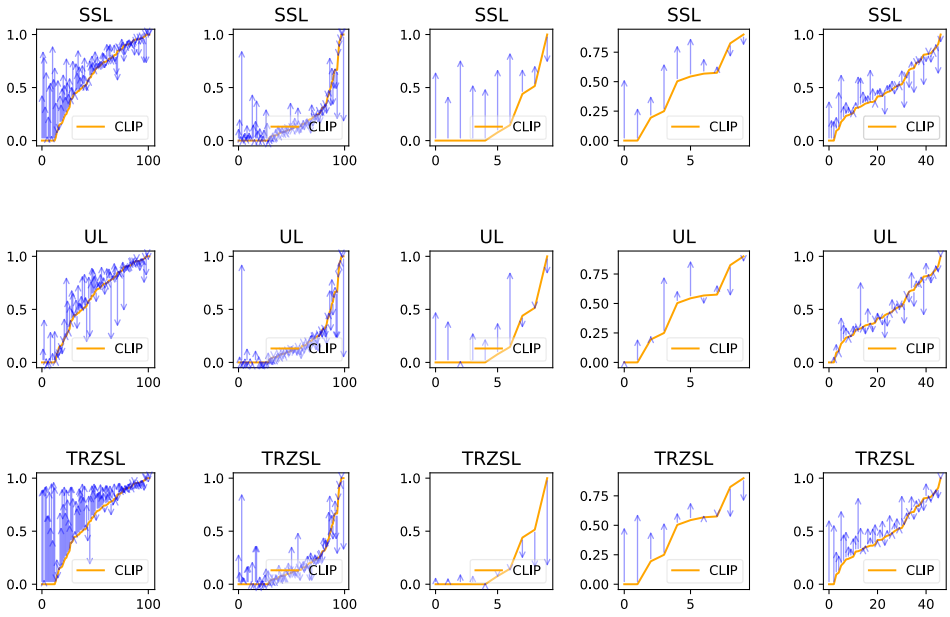
A.4 The Robin Hood effect

The Robin Hood effect on all tasks For each dataset, we provide the per-class accuracy distribution of GRIP compared with CLIP, Figure 8. The Robin Hood effect characterizes all the tasks. We observe that for GRIP the increase in overall accuracy corresponds to consistent improvements in the predictions of initially poor classes. By comparing Figure 7 with Figure 8, we see that GRIP reinforces the Robin Hood effect already visible when using FPL in certain cases.

The importance of good quality pseudolabels to mitigate the Matthew effect in SSL In the SSL setting, we train a logistic regression on top of the visual feature extracted by CLIP's image encoder (ViT-B/32). In Figure 9, we show the per-class accuracy of the final model trained by combining labeled data with either pseudolabels assigned with the conventional scheme (threshold at .95) or 16 CLIP-generated pseudolabels. We compare the two distribution with the per-class accuracy of the model trained solely on the few labeled examples per class (2 instances).

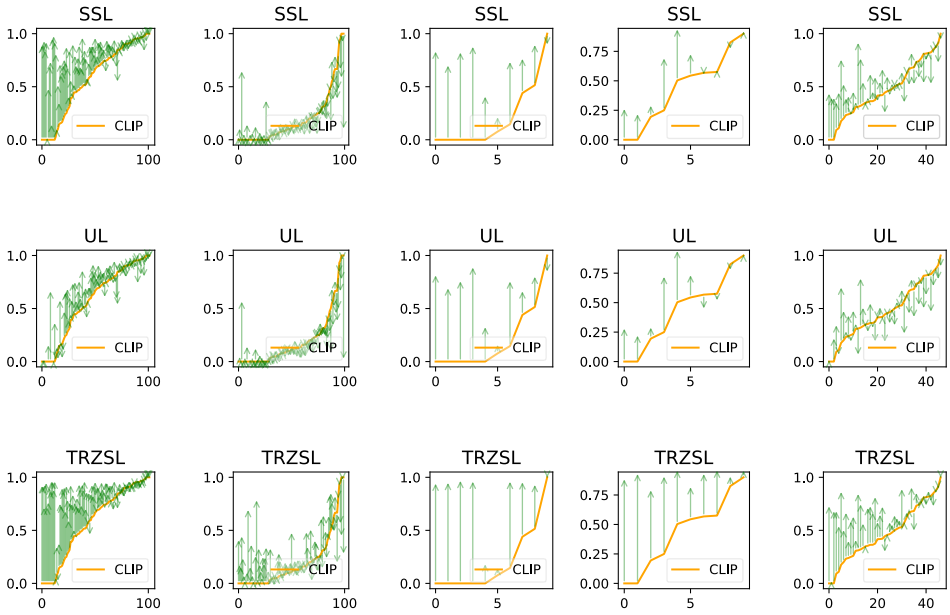
The different impact of prompt tuning and linear probing on the Robin Hood effect We investigate if there is any difference in the Robin Hood effect when adapting CLIP via prompt tuning or linear probing. We train both relying on the iterative training strategy that grows the set of pseudolabels at each iteration by using the top-K scheme (Section 3). We consider the UL setting.

Among the set of target classes, we distinguish between *poor* and *rich* classes. A class is *poor*, if CLIP's accuracy on that class is lower than its overall accuracy on the task. Otherwise, the class is considered *rich*. Table 7 reports the accuracy of the two approaches, and the accuracy on the poor and rich classes, while highlighting the average effect with respect to CLIP. Training with prompt tuning retains more knowledge of the rich classes than linear probing. Prompt tuning reduces the accuracy on the rich classes by on average 0.3 points, while linear probing has an average deterioration of 9.4. Overall, GRIP works better than linear probing. We note that the lower accuracy of linear probing is characterized by a worse ability to correctly predict the rich classes, i.e., "rich get poorer." This is surprising, as we would have expected the errors to concentrate on the poor classes compared to CLIP.



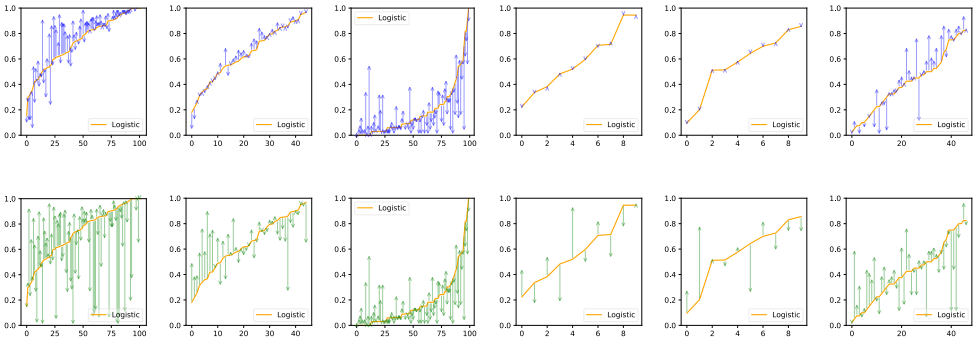
(a) Flowers102 (b) FGVC-Aircraft (c) MNIST (d) EuroSAT (e) DTD

Figure 7: Per-class accuracy of FPL compared to CLIP's per-class accuracy on Flowers102, FGVC-Aircraft, MNIST, EuroSAT, DTD. **X-axis** is the ranked class index, while the **y-axis** is the accuracy.



(a) Flowers102 (b) FGVC-Aircraft (c) MNIST (d) EuroSAT (e) DTD

Figure 8: Per-class accuracy of GRIP compared to CLIP's per-class accuracy on Flowers102, FGVC-Aircraft, MNIST, EuroSAT, and DTD. **X-axis** is the ranked class index, while the **y-axis** is the accuracy.



(a) Flowers102 (b) RESICS45 (c) FGVCA (d) MNIST (e) EuroSAT (f) DTD

Figure 9: Per-class accuracy of a logistic classifier using conventional pseudolabels (first row) and CLIP-based pseudolabels (second row). The solid orange line represents the per-class accuracy of a logistic regression trained on 2-shots per class. **X-axis** is the ranked class index, while the **y-axis** is the accuracy. We present results for Flowers102, RESICS45, FGVC-Aircraft, MNIST, EuroSAT, and DTD, in order.

	Flowers102	RESICS45	FGVC-Aircraft	MNIST	EuroSAT	DTD	Avg. Δ
Linear probe (LP)	41.01	58.79	61.94	50.52	51.37	10.17	-
GRIP	**46.09**	**70.55**	**69.84**	**57.21**	**67.88**	**15.22**	-
Rich CLIP	**67.81**	75.47	85.16	65.26	65.14	**45.93**	-
Rich LP	52.87	69.01	79.55	67.53	50.34	29.12	-
Rich GRIP	56.05	**78.81**	**86.40**	**71.73**	**77.84**	31.95	-
Δ LP	↓ 14.92	↓ 6.47	↓ 5.61	↑ 2.26	↓ 14.79	↓ 16.81	↓ **9.39**
Δ GRIP	↓ 11.76	↑ 3.33	↑ 1.24	↑ 6.46	↑ 12.70	↓ 13.98	↓ 0.33
Poor CLIP	25.63	35.60	27.98	11.10	3.18	5.35	-
Poor LP	26.50	42.77	36.25	28.34	56.76	4.77	-
Poor GRIP	**35.03**	**56.85**	**42.82**	**39.88**	**65.08**	**6.31**	-
Δ LP	↑ 0.87	↑ 7.18	↑ 8.27	↑ 17.24	↑ 53.58	↓ 0.58	↑ 14.43
Δ GRIP	↑ 9.4	↑ 21.26	↑ 14.84	↑ 28.78	↑ 61.9	↑ 0.96	↑ **22.86**

Table 7: For each task we report the overall accuracy of linear probing (LP) and GRIP textual along with the accuracy on *poor* and *rich* classes. Δ `METHOD` is the difference between the accuracy of CLIP and `METHOD`. For an overall evaluation of the difference between linear probing and prompt tuning, we report the average difference of LP and GRIP with respect to CLIP on poor and rich classes.

Split 1

	Flowers102			RESICS45			FGVCAircraft		
Method	S	U	H	S	U	H	S	U	H
CLIP	$64.26_{0.00}$	$62.56_{0.00}$	$63.4_{0.00}$	$54.85_{0.00}$	$54.08_{0.00}$	$54.46_{0.00}$	$16.27_{0.00}$	$19.79_{0.00}$	$17.86_{0.00}$
CoOp	$\mathbf{91.52}_{0.36}$	$48.35_{2.96}$	$63.22_{2.60}$	$\mathbf{84.66}_{1.01}$	$50.73_{3.28}$	$63.37_{2.23}$	$\mathbf{34.18}_{1.56}$	$16.28_{3.69}$	$21.70_{3.45}$
GRIP	$90.31_{0.51}$	$\mathbf{82.57}_{1.26}$	$\mathbf{86.26}_{0.81}$	$82.68_{0.47}$	$\mathbf{79.53}_{0.72}$	$\mathbf{81.07}_{0.37}$	$22.25_{0.07}$	$\mathbf{31.51}_{0.59}$	$\mathbf{26.08}_{0.25}$
Δ CLIP	↑ 26.05	↑ 20.01	↑ 22.86	↑ 27.83	↑ 25.45	↑ 26.61	↑ 5.98	↑ 11.72	↑ 8.22
Δ CoOp	↓ 1.21	↑ 34.22	↑ 23.04	↓ 1.98	↑ 28.8	↑ 17.7	↓ 11.93	↑ 15.23	↑ 4.38

	MNIST			EuroSAT			DTD		
CLIP	$31.74_{0.00}$	$15.43_{0.00}$	$20.77_{0.00}$	$22.33_{0.00}$	$48.3_{0.00}$	$30.54_{0.00}$	$42.5_{0.00}$	$44.44_{0.00}$	$43.45_{0.00}$
CoOp	$\mathbf{94.68}_{5.64}$	$15.43_{7.75}$	$21.15_{12.18}$	$82.91_{8.81}$	$46.02_{9.23}$	$58.64_{5.86}$	$\mathbf{69.67}_{1.17}$	$34.81_{3.44}$	$46.3_{2.92}$
GRIP	$\mathbf{95.13}_{0.11}$	$\mathbf{60.63}_{0.44}$	$\mathbf{74.06}_{0.29}$	$\mathbf{91.75}_{0.53}$	$\mathbf{92.91}_{0.91}$	$\mathbf{92.33}_{0.70}$	$\mathbf{68.26}_{0.69}$	$\mathbf{62.61}_{1.87}$	$\mathbf{65.30}_{1.03}$
Δ CLIP	↑ 63.39	↑ 45.2	↑ 53.29	↑ 69.42	↑ 44.61	↑ 61.79	↑ 25.76	↑ 18.17	↑ 21.85
Δ CoOp	↑ 0.45	↑ 45.2	↑ 52.91	↑ 8.84	↑ 46.89	↑ 33.69	↓ 1.41	↑ 27.8	↑ 19.00

Split 2

	Flowers102			RESICS45			FGVCAircraft		
Method	S	U	H	S	U	H	S	U	H
CLIP	$65.38_{0.00}$	$60.64_{0.00}$	$62.92_{0.00}$	$59.5_{0.00}$	$47.06_{0.00}$	$52.55_{0.00}$	$17.30_{0.00}$	$18.12_{0.00}$	$17.70_{0.00}$
CoOp	$\mathbf{91.8}_{1.32}$	$47.75_{3.86}$	$62.77_{3.31}$	$\mathbf{86.54}_{1.92}$	$48.00_{3.01}$	$61.70_{2.17}$	$\mathbf{33.59}_{4.12}$	$19.57_{1.37}$	$\mathbf{24.63}_{0.63}$
GRIP	$88.84_{0.75}$	$\mathbf{70.93}_{2.08}$	$\mathbf{78.86}_{1.26}$	$84.47_{0.41}$	$\mathbf{84.09}_{1.01}$	$\mathbf{84.28}_{0.73}$	$22.13_{0.24}$	$28.32_{0.33}$	$\mathbf{24.84}_{0.05}$
Δ CLIP	↑ 23.46	↑ 10.29	↑ 15.94	↑ 27.83	↑ 25.45	↑ 26.61	↑ 4.83	↑ 10.20	↑ 7.14
Δ CoOp	↓ 2.96	↑ 23.18	↑ 16.09	↓ 2.07	↑ 36.09	↑ 22.58	↓ 11.46	↑ 8.75	↑ 0.21

	MNIST			EuroSAT			DTD		
CLIP	$15.99_{0.00}$	$39.18_{0.00}$	$22.71_{0.00}$	$32.47_{0.00}$	$33.1_{0.00}$	$32.78_{0.00}$	$45.43_{0.00}$	$39.72_{0.00}$	$42.39_{0.00}$
CoOp	$\mathbf{90.6}_{13.02}$	$18.77_{9.12}$	$30.29_{12.38}$	$86.43_{3.23}$	$47.16_{11.17}$	$60.53_{8.42}$	$\mathbf{70.4}_{1.99}$	$32.53_{4.58}$	$44.42_{4.63}$
GRIP	**95.71**	**97.50**	**96.59**	$\mathbf{91.08}_{0.02}$	$\mathbf{92.02}_{0.98}$	$\mathbf{91.55}_{0.47}$	$66.69_{0.53}$	$\mathbf{56.19}_{1.18}$	$\mathbf{60.99}_{0.69}$
Δ CLIP	↑ 85.12	↑ 50.76	↑ 79.32	↑ 58.61	↑ 58.92	↑ 58.77	↑ 21.26	↑ 16.47	↑ 18.6
Δ CoOp	↑ 6.11	↑ 71.20	↑ 57.19	↑ 4.65	↑ 44.86	↑ 31.02	↓ 3.71	↑ 23.66	↑ 16.57

Split 3

	Flowers102			RESICS45			FGVCAircraft		
Method	S	U	H	S	U	H	S	U	H
CLIP	$68.29_{0.00}$	$57.25_{0.00}$	$62.28_{0.00}$	$56.02_{0.00}$	$52.32_{0.00}$	$54.10_{0.00}$	$17.55_{0.00}$	$17.71_{0.00}$	$17.63_{0.00}$
CoOp	$\mathbf{91.52}_{0.35}$	$48.35_{2.95}$	$63.22_{2.60}$	$\mathbf{87.61}_{2.17}$	$43.64_{4.97}$	$58.14_{4.12}$	$\mathbf{37.77}_{1.92}$	$16.46_{3.23}$	$22.77_{3.09}$
GRIP	$90.09_{0.53}$	$\mathbf{69.00}_{2.44}$	$\mathbf{78.13}_{1.71}$	$85.19_{0.15}$	$\mathbf{75.58}_{3.17}$	$\mathbf{80.07}_{1.79}$	$22.07_{0.23}$	$\mathbf{28.72}_{0.76}$	$\mathbf{24.95}_{0.20}$
Δ CLIP	↑ 21.8	↑ 11.75	↑ 15.85	↑ 29.17	↑ 23.26	↑ 25.97	↑ 4.52	↑ 11.01	↑ 7.32
Δ CoOp	↓ 1.43	↑ 20.65	↑ 14.91	↓ 2.42	↑ 31.94	↑ 21.93	↓ 15.70	↑ 12.26	↑ 2.18

	MNIST			EuroSAT			DTD		
CLIP	$10.59_{0.00}$	$46.74_{0.00}$	$17.27_{0.00}$	$41.47_{0.00}$	$19.60_{0.00}$	$26.62_{0.00}$	$45.52_{0.00}$	$39.58_{0.00}$	$42.34_{0.00}$
CoOp	$89.6_{8.08}$	$26.3_{12.88}$	$39.4_{16.61}$	$79.33_{9.37}$	$43.38_{12.49}$	$55.06_{8.62}$	$\mathbf{70.53}_{3.11}$	$24.94_{5.37}$	$36.63_{5.57}$
GRIP	**95.8**	**96.06**	**95.93**	$\mathbf{90.57}_{0.13}$	$\mathbf{94.25}_{1.10}$	$\mathbf{92.37}_{0.60}$	$67.28_{0.74}$	$\mathbf{58.94}_{2.78}$	$\mathbf{62.81}_{1.75}$
Δ CLIP	↑ 79.81	↑ 56.88	↑ 73.22	↑ 49.1	↑ 74.65	↑ 65.75	↑ 21.76	↑ 19.36	↑ 20.47
Δ CoOp	↑ 5.20	↑ 77.29	↑ 65.64	↑ 11.24	↑ 50.87	↑ 37.31	↓ 3.25	↑ 34.00	↑ 26.18

Table 8: In the TRZSL settings, for each dataset and split, we compare the accuracy of GRIP textual with CLIP zero-shot (ViT-B/32), and CoOp. Results show the accuracy on seen (S) and unseen classes (U), and the harmonic mean (H). We average the accuracy on 5 seeds and report the standard deviation. Δ `METHOD` is the difference between the accuracy of GRIP and `METHOD`.

Split 1	Seen classes (S)	Unseen classes (U)
Flowers102	canna lily, petunia, silverbush, prince of wales feathers, pincushion flower, bird of paradise, frangipani, hard-leaved pocket orchid, bearded iris, passion flower, tiger lily, lenten rose, cape flower, air plant, mexican petunia, common dandelion, magnolia, foxglove, hibiscus, camellia, orange dahlia, clematis, anthurium, bougainvillea, ruby-lipped cattleya, stemless gentian, oxeye daisy, spring crocus, king protea, cyclamen, fritillary, californian poppy, wild pansy, desert-rose, sunflower, rose, grape hyacinth, pink primrose, red ginger, corn poppy, watercress, colt's foot, blanket flower, monkshood, morning glory, siam tulip, barbeton daisy, bolero deep blue, carnation, tree poppy, globe thistle, english marigold, primula, wallflower, blackberry lily, fire lily, love in the mist, moon orchid, sweet pea, mallow, pelargonium, mexican aster, poinsettia	canterbury bells, snapdragon, spear thistle, yellow iris, globe flower, purple coneflower, peruvian lily, balloon flower, giant white arum lily, artichoke, sweet william, garden phlox, alpine sea holly, great masterwort, daffodil, sword lily, marigold, buttercup, bishop of llandaff, gaura, geranium, pink and yellow dahlia, cautleya spicata, japanese anemone, black-eyed susan, osteospermum, windflower, gazania, azalea, water lily, thorn apple, lotus, toad lily, columbine, tree mallow, hippeastrum, bee balm, bromelia, trumpet creeper
RESICS45	beach, palace, roundabout, railway station, railway, thermal power station, river, airplane, island, bridge, basketball court, desert, runway, ground track field, sea ice, sparse residential, cloud, dense residential, wetland, mountain, meadow, baseball diamond, parking lot, storage tank, tennis court, commercial area, mobile home park	airport, ship, snowberg, chaparral, church, circular farmland, stadium, terrace, forest, freeway, golf course, harbor, industrial area, intersection, lake, medium residential, overpass, rectangular farmland
FGVC-Aircraft	Tu-134, Spitfire, Challenger 600, 737-700, F-A-18, E-170, 727-200, A300B4, Falcon 2000, DR-400, MD-87, CRJ-700 ERJ 145, Falcon 900, MD-80, DC-10, Il-76, Global Express, Gulfstream IV, Saab 340, Yak-42, CRJ-900, L-1011, A330-200, A321, 747-300, DC-3, A310, ATR-42, CRJ-200, Hawk T1, Fokker 100, ATR-72, PA-28, A319, 707-320, A318, A320, BAE-125, 747-200, ERJ 135, 737-800, SR-20, BAE 146-300, Beechcraft 1900, Cessna 172, A340-300, EMB-120, 737-900, 737-400, Cessna 208, MD-90, 777-300, A340-600, 737-600, 737-300, DHC-1, DC-6, A380, C-47, 767-200, BAE 146-200	737-200, 737-500, 747-100, 747-400, 757-200, 757-300, 767-300, 767-400, 777-200, A330-300, A340-200, A340-500, An-12, Boeing 717, C-130, Cessna 525, Cessna 560, DC-8, DC-9-30, DH-82, DHC-6, DHC-8-100, DHC-8-300, Dornier 328, E-190, E-195, Embraer Legacy 600, Eurofighter Typhoon, F-16A-B, Fokker 50, Fokker 70, Gulfstream V, MD-11, Metroliner, Model B200, Saab 2000, Tornado, Tu-154
MNIST	4, 2, 9, 3, 0, 5	8, 1, 6, 7
EuroSAT	industrial buildings or commercial buildings, brushland or shrubland, lake or sea, highway or road, annual crop land, pasture land	river, forest, permanent crop land, residential buildings or homes or apartments
DTD	knitted, pitted, studded, bumpy, spiralled, scaly, polka-dotted, veined, wrinkled, banded, flecked, stained, chequered, sprinkled, bubbly, grid, lined, crystalline, fibrous, meshed, zigzagged, pleated, braided, perforated, potholed, waffled, dotted, matted, gauzy	blotchy, smeared, cobwebbed, cracked, crosshatched, stratified, striped, swirly, woven, freckled, frilly, grooved, honeycombed, interlaced, lacelike, marbled, paisley, porous
Split 2		
Flowers102	prince of wales feathers, air plant, canterbury bells, bishop of llandaff, bee balm, desert-rose, purple coneflower, spring crocus, pelargonium, windflower, sunflower, bougainvillea, rose, spear thistle, bird of paradise, carnation, fritillary, grape hyacinth, mexican aster, monkshood, poinsettia, black-eyed susan, sweet pea, anthurium, wallflower, oxeye daisy, moon orchid, blackberry lily, hibiscus, frangipani , cautleya spicata, camellia, canna lily, passion flower, wild pansy, stemless gentian, balloon flower, gaura, thorn apple, morning glory, hard-leaved pocket orchid, japanese anemone, sword lily, daffodil, english marigold, globe flower, peruvian lily, barbeton daisy, siam tulip, tiger lily, foxglove, pink and yellow dahlia, pink primrose, alpine sea holly, artichoke, petunia, colt's foot, ruby-lipped cattleya, red ginger, primula, snapdragon, garden phlox, mexican petunia	globe thistle, king protea, yellow iris, giant white arum lily, fire lily, pincushion flower, corn poppy, sweet william, love in the mist, cape flower, great masterwort, lenten rose, bolero deep blue, marigold, buttercup, common dandelion, geranium, orange dahlia, silverbush, californian poppy, osteospermum, bearded iris, tree poppy, gazania, azalea, water lily, lotus, toad lily, clematis, columbine, tree mallow , magnolia, cyclamen, watercress, hippeastrum, mallow, bromelia, blanket flower, trumpet creeper
RESICS45	railway station, snowberg, palace, beach, commercial area, mountain, parking lot, dense residential, sparse residential, rectangular farmland, railway, island, tennis court, baseball diamond, thermal power station, industrial area, golf course, meadow, ground track field, storage tank, circular farmland, forest, bridge, harbor, river, freeway, sea ice	airplane, airport, roundabout, basketball court, runway, ship, chaparral, church, stadium, cloud, terrace, desert, wetland, intersection, lake, medium residential, mobile home park, overpass
FGVC-Aircraft	A321, MD-80, 737-200, DC-8, Falcon 900, Saab 340, 767-200, F-A-18, DC-6, SR-20, DC-3, Saab 2000, Fokker 70, 747-400, 737-700, A340-300, A310, A319, A380, 737-800, C-47, Dornier 328, 737-300, Eurofighter Typhoon, Cessna 208, Challenger 600, 737-600, Yak-42, Hawk T1, Fokker 100, DHC-8-100, Gulfstream IV, Model B200, Embraer Legacy 600, CRJ-900, A330-200, 767-400, DC-9-30, DR-400, Falcon 2000, 727-200, DHC-8-300, C-130, Boeing 717, 737-400, 757-300, 767-300, Beechcraft 1900, BAE 146-300, 737-500, PA-28, DHC-6, 707-320, An-12, A330-300, CRJ-700, 747-200, ATR-42, A318, DC-10, 747-100, A340-500	737-900, 747-300, 757-200, 777-200, 777-300, A300B4, A320, A340-200, A340-600, ATR-72, BAE 146-200, BAE-125, Cessna 172, Cessna 525, Cessna 560, CRJ-200, DH-82, DHC-1, E-170, E-190, E-195, EMB-120, ERJ 135, ERJ 145, F-16A-B, Fokker 50, Global Express, Gulfstream V, Il-76, L-1011, MD-11, MD-87, MD-90, Metroliner, Spitfire, Tornado, Tu-134, Tu-154
MNIST	2, 8, 4, 9, 1, 6	0, 3, 5, 7
EuroSAT	brushland or shrubland, river, industrial buildings or commercial buildings, lake or sea, forest, permanent crop land	annual crop land, highway or road, pasture land, residential buildings or homes or apartments
DTD	pitted, scaly, polka-dotted, bumpy, honeycombed, fibrous, veined, porous, lined, dotted, perforated, potholed, pleated, waffled, braided, wrinkled, paisley, gauzy, meshed, grid, studded, knitted, swirly, crosshatched, freckled, chequered, grooved, smeared, frilly	banded, blotchy, bubbly, spiralled, sprinkled, cobwebbed, cracked, stained, crystalline, stratified, striped, flecked, woven, zigzagged, interlaced, lacelike, marbled, matted
Split 3		
Flowers102	oxeye daisy, canterbury bells, clematis, siam tulip, cape flower, black-eyed susan, air plant, californian poppy, globe thistle, giant white arum lily, cyclamen, snapdragon, frangipani, buttercup, common dandelion, hippeastrum, columbine, spring crocus, bolero deep blue, spear thistle, barbeton daisy, poinsettia, peruvian lily, alpine sea holly, artichoke, sunflower, tiger lily, toad lily, magnolia, lenten rose, great masterwort, camellia, mallow, morning glory, lotus, sweet william, thorn apple, carnation, daffodil, corn poppy, cautleya spicata, marigold, hibiscus, tree poppy, balloon flower, osteospermum, english marigold, king protea, azalea, foxglove, watercress, blackberry lily, bearded iris, monkshood, mexican aster, orange dahlia, water lily, mexican petunia, sweet pea, pink primrose, primula, silverbush, pincushion flower	hard-leaved pocket orchid, moon orchid, bird of paradise , colt's foot, yellow iris, globe flower, purple coneflower, fire lily, fritillary, red ginger, grape hyacinth, prince of wales feathers, stemless gentian, garden phlox, love in the mist, ruby-lipped cattleya, sword lily, wallflower, petunia, wild pansy, pelargonium, bishop of llandaff, gaura, geranium, pink and yellow dahlia, japanese anemone, windflower, gazania, rose, passion flower, anthurium, desert-rose, tree mallow, canna lily, bee balm, bougainvillea, bromelia, blanket flower, trumpet creeper
RESICS45	railway, parking lot, wetland, meadow, harbor, island, mobile home park, storage tank, industrial area, bridge, baseball diamond, sea ice, runway, airplane, thermal power station, circular farmland, basketball court, roundabout, commercial area, railway station, terrace, forest, rectangular farmland, lake, medium residential, snowberg, river	airport, beach, ship, chaparral, church, sparse residential, cloud, stadium, dense residential, desert, tennis court, freeway, golf course, ground track field, intersection, mountain, overpass, palace
FGVC-Aircraft	An-12, 737-200, F-16A-B, BAE 146-200, MD-80, E-170, Gulfstream IV, DR-400, 737-900, 777-200, Boeing 717, 747-100, Saab 340, Cessna 525, Challenger 600, MD-90, DHC-8-100, Cessna 172, C-47, 747-400, BAE-125, MD-11, 767-300, Cessna 560, A330-300, E-195, 737-500, Fokker 50, ATR-72, BAE 146-300, Fokker 70, Falcon 900, Falcon 2000, Spitfire, A340-200, DC-3, A340-300, Beechcraft 1900, A320, Hawk T1, E-190, Gulfstream V, Tu-134, 767-400, CRJ-200, 737-400, 747-300, Eurofighter Typhoon, PA-28, MD-87, Yak-42, DHC-1, 737-800, A380, Model B200, ERJ 135, SR-20, 737-300, 707-320, DC-10, Dornier 328, A300B4	727-200, 737-600, 737-700, 747-200, 757-200, 757-300, 767-200, 777-300, A310, A318, A319, A321, A330-200, A340-500, A340-600, ATR-42, C-130, Cessna 208, CRJ-700, CRJ-900, DC-6, DC-8, DC-9-30, DH-82, DHC-6, DHC-8-300, EMB-120, Embraer Legacy 600, ERJ 145, F-A-18, Fokker 100, Global Express, Il-76, L-1011, Metroliner, Saab 2000, Tornado, Tu-154
MNIST	8, 3, 5, 6, 1, 7	0, 9, 2, 4
EuroSAT	river, highway or road, pasture land, permanent crop land, forest, residential buildings or homes or apartments	annual crop land, lake or sea, brushland or shrubland, industrial buildings or commercial buildings
DTD	pitted, pleated, polka-dotted, sprinkled, grooved, knitted, matted, wrinkled, honeycombed, chequered, braided, zigzagged, spiralled, banded, waffled, crosshatched, bubbly, smeared, dotted, porous, woven, freckled, lined, potholed, lacelike, marbled, stratified, scaly, studded	blotchy, bumpy, stained, cobwebbed, cracked, striped, crystalline, swirly, fibrous, flecked, veined, frilly, gauzy, grid, interlaced, meshed, paisley, perforated

Table 9: For each dataset, we report the class names of seen and unseen classes in each of the splits used for TRZSL.