Appendix

A. VP Preliminary

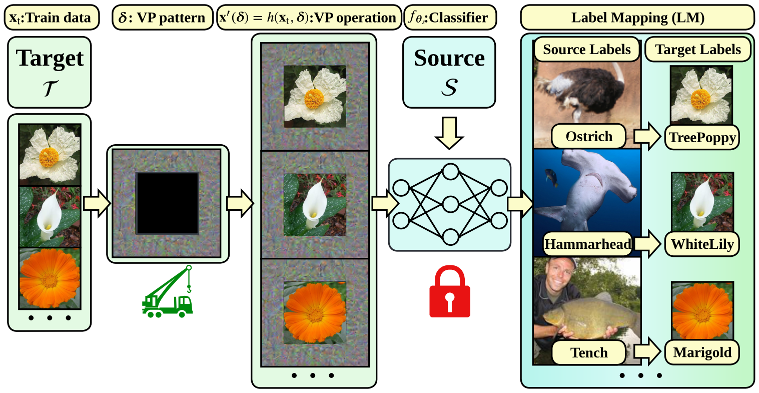

Fig. A1. Overview of a vanilla VP pipeline. VP consists of (1) input prompting operation to incorporate prompt pattern into target data; (2) prompt generation by optimizing prompt pattern with the frozen source classifier; (3) label mapping for the upstream classifier to execute the downstream task.

Fig. A1 shows the pipeline of the vanilla VP method. The method has three main procedures: (1) input prompting operation to inject each image of the target dataset into the VP pattern; (2) prompt generation to optimize the VP pattern with the fixed, pre-trained source classifier; (3) pre-defined label mapping from the source data labels to the target data labels.

B. Algorithm Overview of Our Proposal: ILM-VP

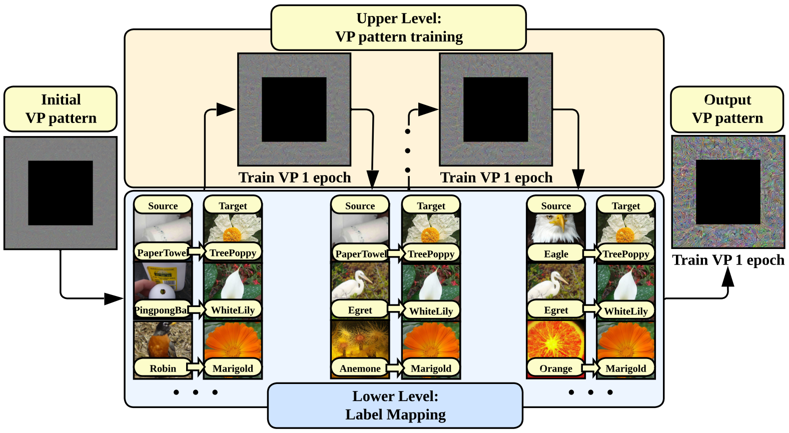

Fig. A2. Algorithm overview of ILM-VP. Alternating optimization is iteratively executed between the upper-level prompt generation and the lower-level LM. This BLO process can progressively improve both downstream task accuracy and LM interpretability.

Fig. A2 shows the bi-level optimization pipeline of our proposal ILM-VP. The method has two levels. (1) Lower level: Given the VP pattern from the previous epoch, use FLM technique to update the label mapping for each target class. (2) Upper level: Given the label mapping, call SGD to update the VP pattern.

C. Explanation by Examples

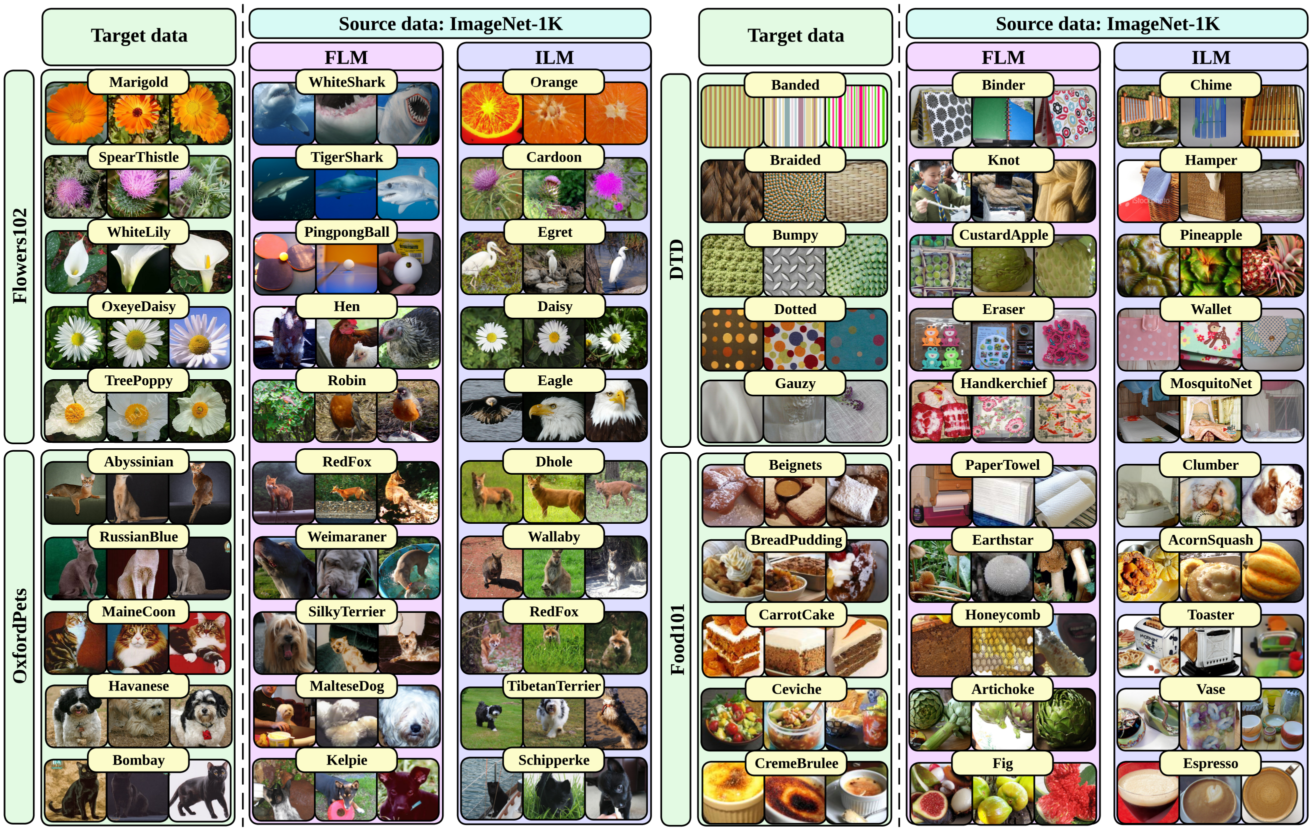

Fig. A3. LM method comparison: FLM vs ILM, by visualization of the target dataset (1) FLowers102, (2) OxfordPets, (3) DTD, and (4) Food101 with the source dataset ImageNet-1K using pretrained ResNet-18. ILM consistently finds more interpretable LM than FLM, in terms of colors, scenes, shapes, and textures. The four datasets are chosen due to the resolution for visualization and the accuracy improvement by our method compared to the prior art.

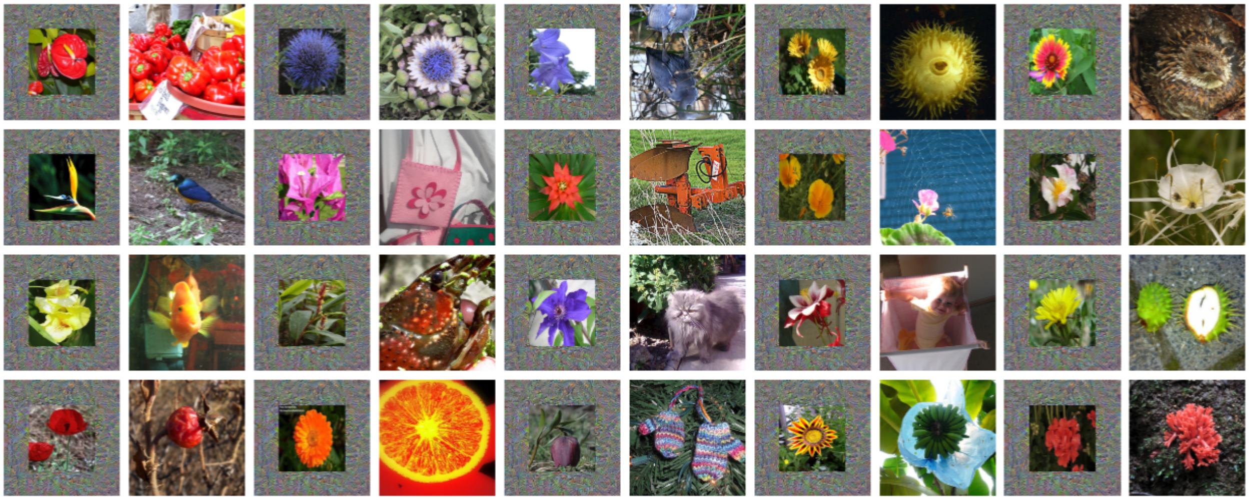

Fig. A4. More EBE examples in Flowers102. Every two pictures refer to one target class image and one corresponding source class image.

Implementation details. Technically, given a test sample, EBE (explanation by example) finds the image from the training set which has the highest cosine similarity with the test datapoint. The cosine similarity is evaluated in the feature space, which is typically chosen as the activations of the penultimate layer of the neural network. We use a target task image integrated with the learned VP as the input sample and search for the most similar training ImageNet samples in the mapped source class. The top-3 most similar source examples are used to explain the prediction using the pre-trained source model over the VP-injected test sample.

Discussion. In VP, the source network preserves the full information of the source training dataset. Thus, EBE in VP makes a bridge between the source dataset and the target dataset, which enables us to better understand the prediction ability of VP. Fig. A3 shows the EBE results of FLM-VP and ILM-VP. Here we can see that ILM-VP discovers more semantically-similar target-source pairs. For example, 'Oxeye Daisy' from Flowers102 is directly correlated with 'Daisy' using ILM rather than 'Hen' using FLM; 'Bombay' from OxfordPets is correlated with a black dog feed 'Schipperke' rather than a brown one 'Kelpie'; 'Banded' from DTD shares similar string with 'Chime' using ILM rather than 'Binder' selected by FLM; 'Bread Pudding' from Food101 is mapped from 'Acorn Squash' with similar textures and food property using ILM instead of 'Earthstar' using FLM. We argue that this high interpretability enabled by our proposed ILM-VP method contributes to the analysis of transfer learning. Recent work on transfer learning has found that the source dataset has class-wise positive or negative influence in downstream tasks [56]. Not surprisingly, they found a positive source class that shares some similar semantic features with one class in the target set. This is also revealed by our method (*e.g.*, 'Orange' mapped to 'Marigold'). Since the method used in [56] is computationally heavy, our method could be a light alternative to identify the most beneficial source class for each downstream class. We show more EBE results in Fig. A4.

D. VP Using Vision Models: Additional Experiments and Details

Datasets. We focus on downstream image classification tasks in the target domain. To highlight a few of them, OxfordPets shares many same labels as ImageNet (*e.g.*, beagle, boxer) while some datasets like GTSRB contain image classes quite different from ImageNet. CIFAR10 possesses a large training set (50000 images) yet a small label space (10 classes) while SUN397 has a rather small training set (15888 images) yet a huge label space (397 classes). In addition, image classes in StanfordCars share quite similar features (*e.g.*, different models and years of cars) while DTD contains rather different data features. Further, ABIDE is a medical dataset that converts the original 1D numerical input sequences to image-alike data formats. We show each dataset's attributes in Tab. A1.

Implementation details. (Code link is available in the zipped supplementary file.) We have RLM-VP (random label mapping based visual prompt), FLM-VP (frequency label mapping based visual prompt), LP (linear probing) and FF (full finetuning) as our baselines. We follow the same implementation in [1] to achieve the frequency-based label mapping (FLM). In ILM-VP, we execute FLM at the beginning of each epoch. To train VP, we choose Adam optimizer and use a learning rate of 0.01 determined by a multi-step decaying scheduler. The total number of trianing epochs is 200. For datasets containing images of different resolutions (*e.g.,* UCF101, Flowers102), we resize target task images into a fixed resolution. Target task resolutions and batch sizes of all datasets are shown in Tab. A1. For LP and FF, we resize the target task images into 224×224. In LP, we use the Adam optimizer, a multi-step decaying scheduler and a learning rate of 0.1. In FF, we use the Adam optimizer, a cosine-annealing scheduler and a learning rate of 0.01. In the FF experiment, we add data augmentation (*i.e.,* Random Crop, Random Flip) and weight decay (with the magnitude of 5×10^{-4}).

Dataset	Train Size	Test Size	Class Number	Batch Size	Rescaled Resolution
Flowers102	4093	2463	102	256	128×128
DTD	2820	1692	47	64	128×128
UCF101	7639	3783	101	256	128×128
Food101	50500	30300	101	256	128×128
SVHN	73257	26032	10	256	32×32
GTSRB	39209	12630	43	256	32×32
EuroSAT	13500	8100	10	256	128×128
OxfordPets	2944	3669	37	64	128×128
StanfordCars	6509	8041	196	256	128×128
SUN397	15888	19850	397	256	128×128
CIFAR10	50000	10000	10	256	32×32
CIFAR100	50000	10000	100	256	32×32
ABIDE	931	104	2	64	200×200

Tab. A1. Dataset attributes and training configs through 13 target image-classification datasets.

Architecture	ILM-VP	FLM-VP	LP	FF
ResNet18	20	14	15	33
ResNet50	26	17	17	49
ResNeXt-101-32x8d	62	45	47	104

Tab. A2. V100 run-time on Flowers102 in minutes.

E. VP Using CLIP: Additional Experiment Details and Results

Implementation details. Fig. A5 shows the difference between the prior art (upper one) and our proposal (lower one). In the prior art, 'This is a photo of a {}' is the only context prompt selected for all different target labels, then correlated with the image encoding to accomplish the target image classification task. In contrast, our proposal considers all 81 different text prompts introduced in the original CLIP setting. We incorporates the label mapping technique to map virtual source labels (given by the combinations of context prompts and target classes) to target labels. We strictly follow the implementation detailed in [2] for the VP+TP baseline. The only difference is that we train 200 epochs instead of 1000 epochs. We observed that the performance is similar and 200 epochs' training is much more efficient. For our VP+TP+LM method, we first create multiple virtual source labels and then do FLM at the beginning of each epoch. The rest setting is the same as the baseline.

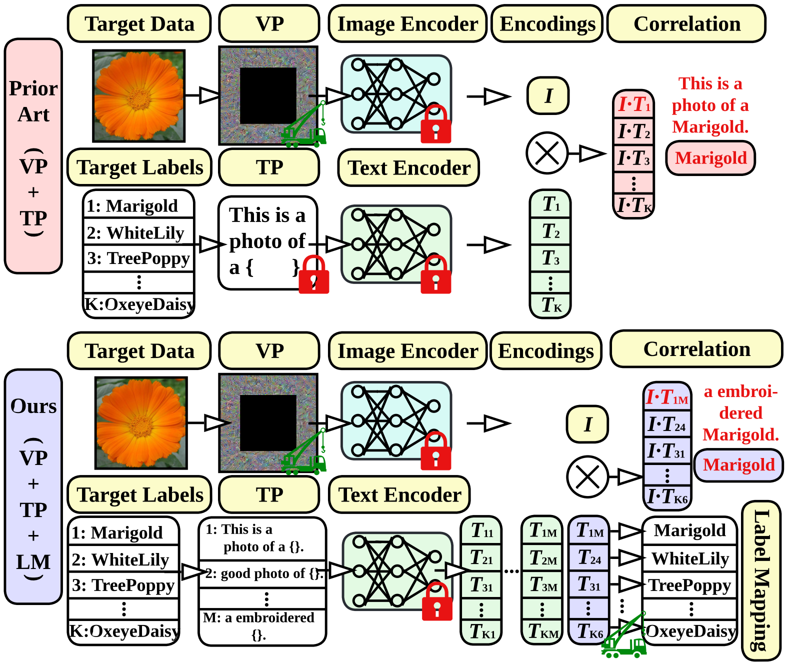

Fig. A5. Overview of CLIP-based prompt learning: our proposal vs. SOTA method. Our proposal incorporates label mapping into the text prompt selection of CLIP.

Interpretability merit of proposed VP+TP+LM method. We show in Fig. A6 the interpretability of prompts found by our method. For example, in Flowers102, 'Tiger Lily' chooses 'A close-up photo of Tiger Lily' as its text prompt. This is reasonable as the dataset consists most of close-up photos of flowers. In CIFAR10, the "Airplane" class chooses "A pixelated photo of Airplane" for the same reason. In addition, there also exist cross-class mappings. For example, 'Hard Leaved Pocket Orchid' corresponds to 'A photo of the large Moon Orchid'. We can see that both 'Hard Leaved Pocket Orchid' and 'Moon Orchid' belong to 'Orchid' and the former one is larger in size.

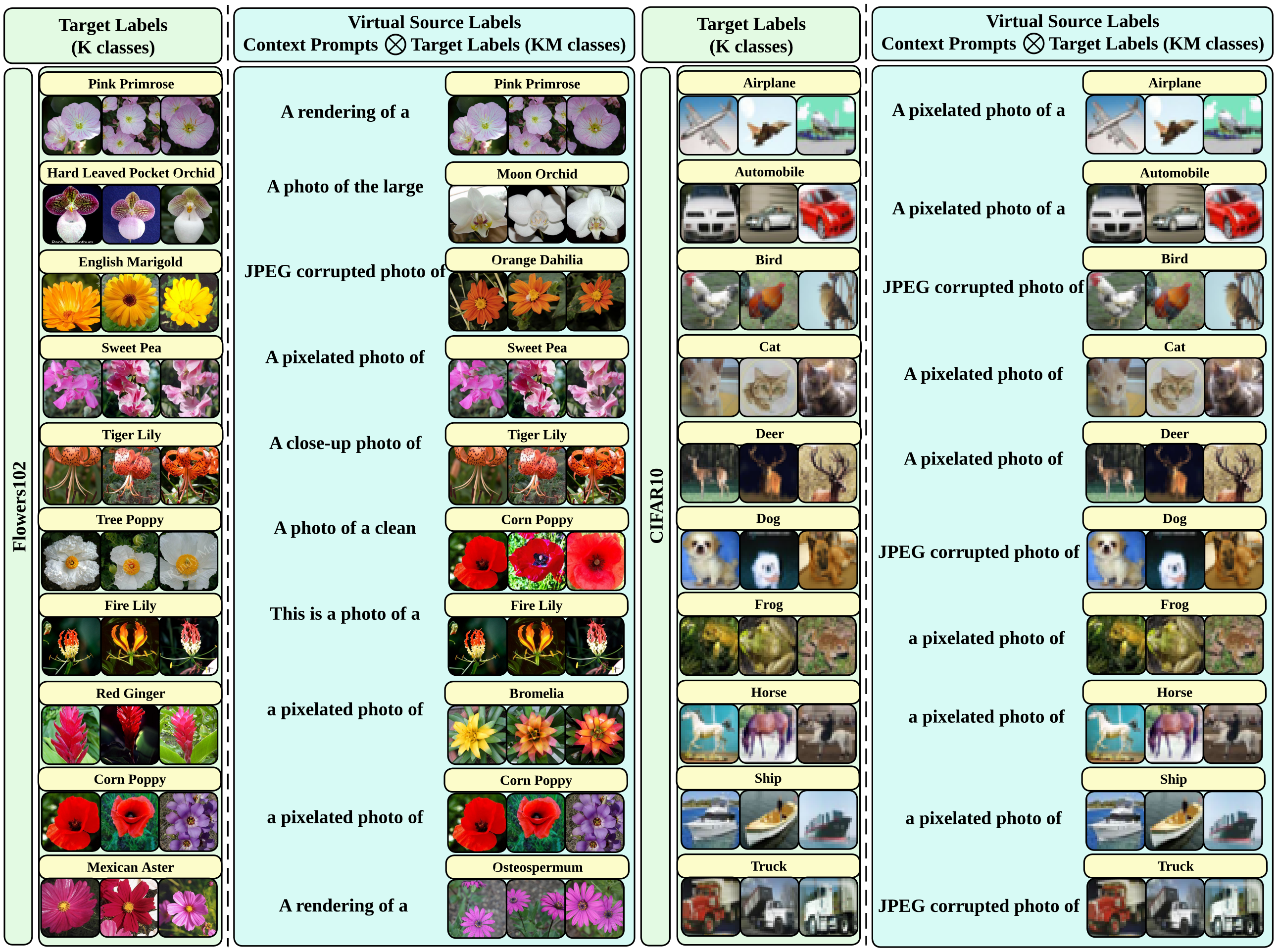

Fig. A6. Label mapping results of our proposed VP method for CLIP. The presented two datasets ‘Flowers102’ and ‘CIFAR10’ shows significant interpretability when using label mapping.