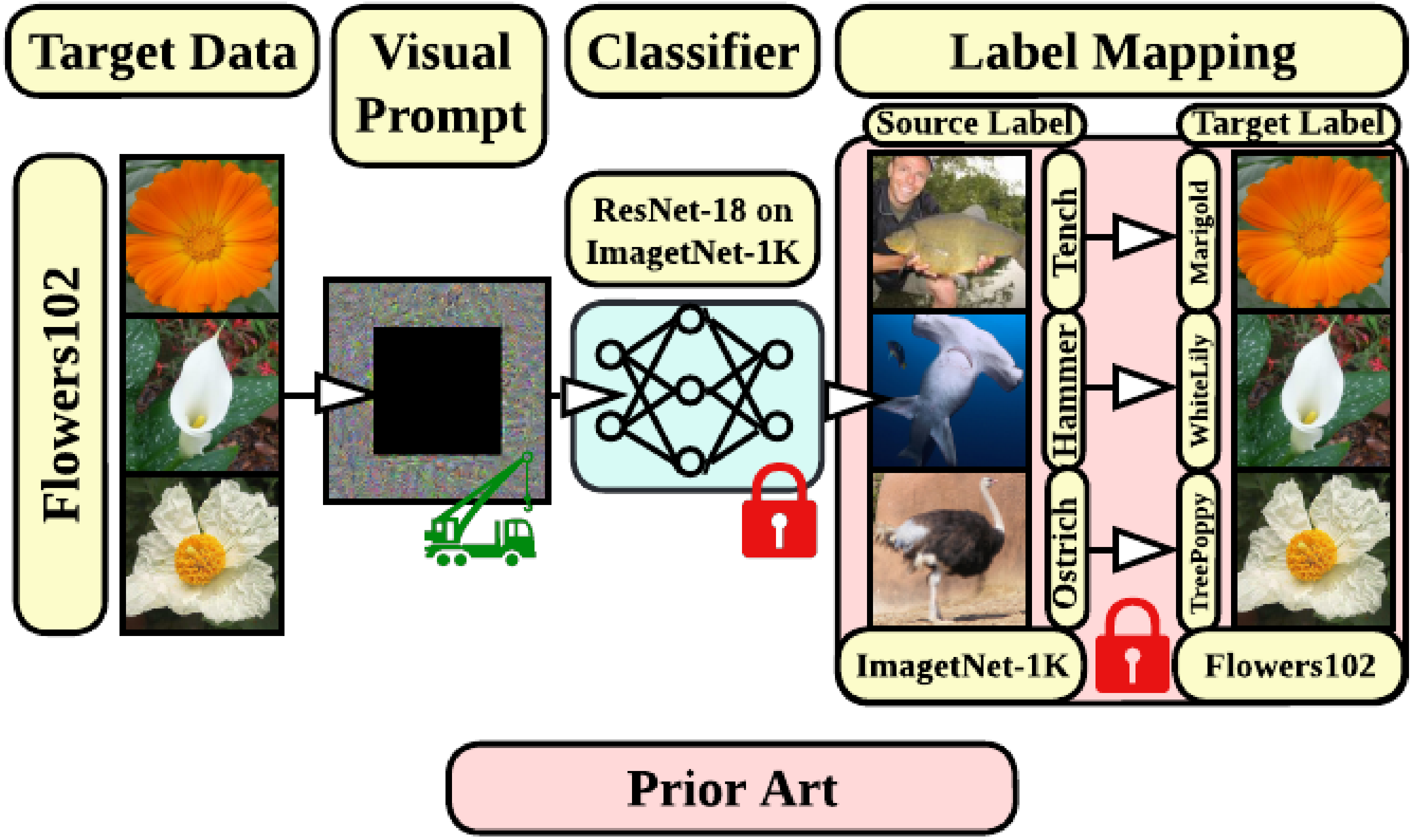
Understanding and Improving Visual Prompting: A Label-Mapping Perspective

Aochuan Chen (MSU), Yuguang Yao (MSU), Pin-Yu Chen (IBM Research), Yihua Zhang (MSU), Sijia Liu (MSU)

Visual Prompting for Transfer Learning



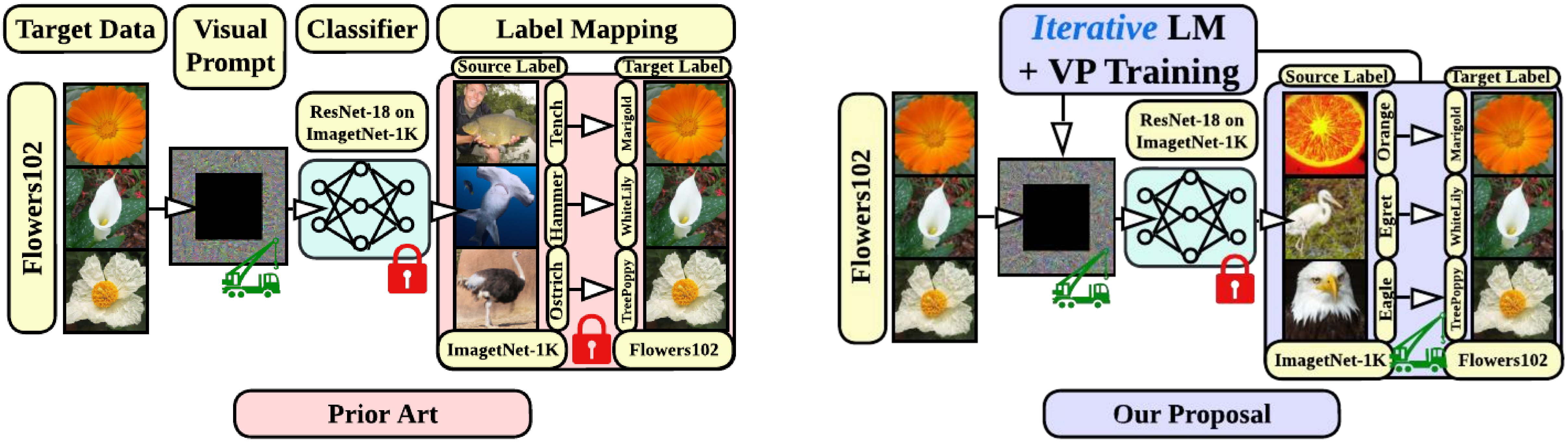
Label Mapping

- Mapping from source classes to downstream classes.
- Existing label mapping methods seem **ruleless**.
- A key building block for visual prompting.

Main Research Question

Given the source model, how to build a mapping from the source label space to the target label space so that the model's prediction directs to the correct target label?

Our Proposal: BLO based ILM-VP



Algorithm 2 Bi-level optimization based ILM-VP algorithm

1: **Initialize:** Given target training set $\mathcal{T}_{\mathrm{tr}}$, pre-trained model $f_{\boldsymbol{\theta}_{\mathrm{s}}}$, prompt pattern initialization $\boldsymbol{\delta}_0$, and upper-level learning rate λ for SGD
2: **for** Epoch $n = 0, 1, \ldots,$ **do**
3: **Lower-level label mapping**: Given $\boldsymbol{\delta}_{n-1}$, call LM for each target class y_{t} in $\mathcal{T}_{\mathrm{tr}}$
4: **Upper-level prompt learning**: Given LM, call SGD to update prompt $\boldsymbol{\delta}_n \leftarrow \boldsymbol{\delta}_{n-1}$
5: **end for**

Convergence of ILM-VP

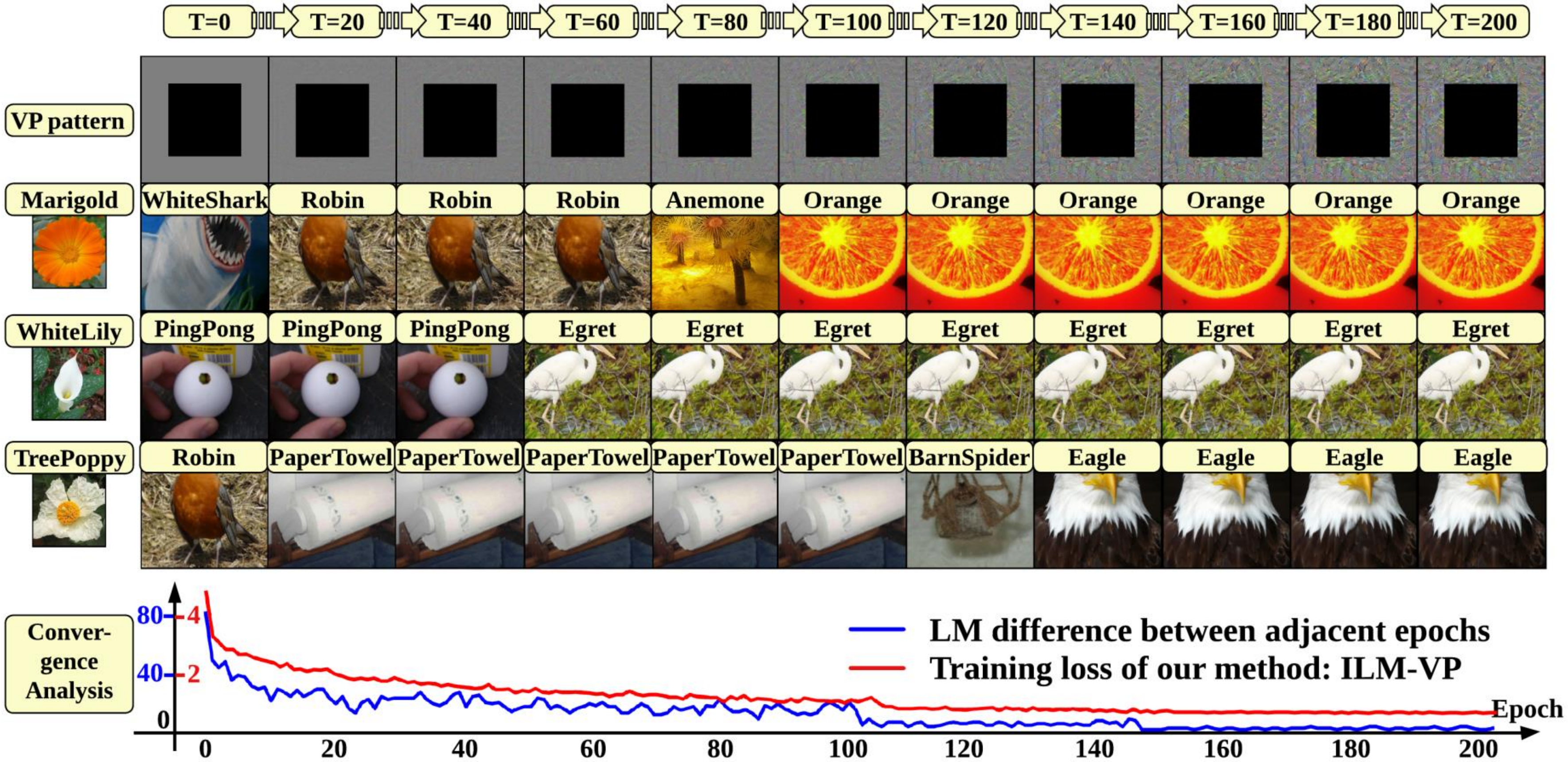


Figure 1. ILM-VP training dynamics from epoch 0 to 200. Rows show: (1) VP pattern vs. epoch number; (2-4) Learned source label mapping with respect to target label 'Marigold', 'White Lily', and 'Tree Poppy', together with explanation-by-example-identified source training examples to explain each re-purposed target label; (5) Convergence of training loss and LM difference between adjacent epochs measured by Hamming distance.

Accuracy Improvements

Source Model	ResNet-18 (ImageNet-1K)					ResNet-50 (ImageNet-1K)				ResNeXt-101-32x8d (Instagram)			
Method	Ours	Prompt baseline		Finetuning		Ours	Prompt base.	Finetuning		Ours	Prompt base.	Finetuning	
	ILM-VP	RLM-VP	FLM-VP	LP	FF	ILM-VP	FLM-VP	LP	FF	ILM-VP	FLM-VP	LP	FF
Parameter Size	0.05M	0.05M	0.05M	0.51M	11.7M	0.05M	0.05M	0.51M	25.6M	0.05M	0.05M	0.51M	88.8M
Flowers102	**27.9**±0.7	11.0±0.5	20.0±0.3	88.0±0.5	97.1±0.7	**24.6**±0.6	20.3±0.3	90.9±0.4	97.9±0.7	**27.9**±0.3	22.5±0.5	89.1±0.2	99.2±0.5
DTD	**35.3**±0.9	16.3±0.7	32.4±0.5	60.0±0.6	65.5±0.9	**40.5**±0.5	36.9±0.8	67.6±0.3	69.7±0.9	**41.4**±0.7	40.3±0.5	69.7±0.2	69.1±1.0
UCF101	**23.9**±0.5	6.6±0.4	18.9±0.5	63.2±0.8	73.0±0.6	**34.6**±0.2	33.9±0.4	70.8±0.3	78.0±0.8	**43.1**±0.8	41.9±0.6	76.9±0.5	79.1±0.7
Food101	**14.8**±0.2	3.8±0.3	12.8±0.1	50.6±0.3	75.4±0.8	**17.0**±0.3	15.3±0.2	57.6±0.5	80.3±0.9	**23.0**±0.4	20.5±0.5	76.0±0.4	82.5±0.3
GTSRB	**52.0**±1.2	46.1±1.3	45.5±1.0	77.4±1.2	98.0±0.3	**52.5**±1.4	47.6±1.1	77.8±0.7	97.6±1.0	**59.9**±1.0	56.2±0.6	73.5±0.7	97.6±0.9
EuroSAT	**85.2**±0.6	82.4±0.4	83.8±0.2	93.8±0.3	98.8±0.5	83.6±0.7	**84.8**±0.3	95.7±0.2	98.9±0.6	86.2±0.8	**87.8**±0.4	93.4±0.3	98.9±0.7
OxfordPets	**65.4**±0.7	9.3±0.4	62.9±0.1	87.2±0.6	87.8±0.5	76.2±0.6	**76.4**±0.2	90.4±0.3	91.9±0.4	**78.9**±0.8	76.8±0.6	93.6±0.4	90.1±0.9
StanfordCars	**4.5**±0.1	0.9±0.1	2.7±0.1	33.8±0.2	81.0±0.1	**4.7**±0.2	4.2±0.3	40.6±0.1	86.4±0.3	**7.0**±0.2	4.6±0.1	64.7±0.1	92.5±0.2
SUN397	**13.0**±0.2	1.0±0.1	10.4±0.1	46.1±0.2	53.2±0.2	**20.3**±0.2	19.8±0.1	53.5±0.1	59.0±0.1	**23.7**±0.2	21.6±0.3	62.3±0.1	61.0±0.2
CIFAR10	65.5±0.1	63.0±0.1	**65.7**±0.6	85.9±0.5	96.5±0.4	**76.6**±0.3	74.8±0.5	90.1±0.1	96.6±0.2	**81.7**±0.3	80.3±0.3	94.1±0.1	97.1±0.1
CIFAR100	**24.8**±0.1	12.9±0.1	18.1±0.2	63.3±0.8	82.5±1.2	**38.9**±0.3	32.0±0.4	70.7±0.7	83.4±0.9	**45.9**±0.2	39.7±0.2	76.2±0.9	84.6±1.2
SVHN	**75.2**±0.2	73.5±0.3	73.1±0.2	65.0±0.2	96.5±0.3	**75.8**±0.4	75.6±0.2	63.5±0.2	96.9±0.3	**81.4**±0.1	79.0±0.5	51.0±0.2	97.1±0.3
ABIDE	**76.9**±2.1	74.0±2.2	73.1±1.6	65.4±3.8	60.6±4.2	63.5±2.2	**64.4**±3.4	55.8±2.6	70.2±2.5	**67.3**±2.6	65.7±3.4	54.8±3.4	73.1±4.2

Table 1. (Main Results) Performance overview of ILM-VP, prompt baseline methods (RLM-VP and FLM-VP), and finetuning methods (LP and FF) over 13 target image classification datasets using 3 pretrained source models.

Target Dataset Analysis

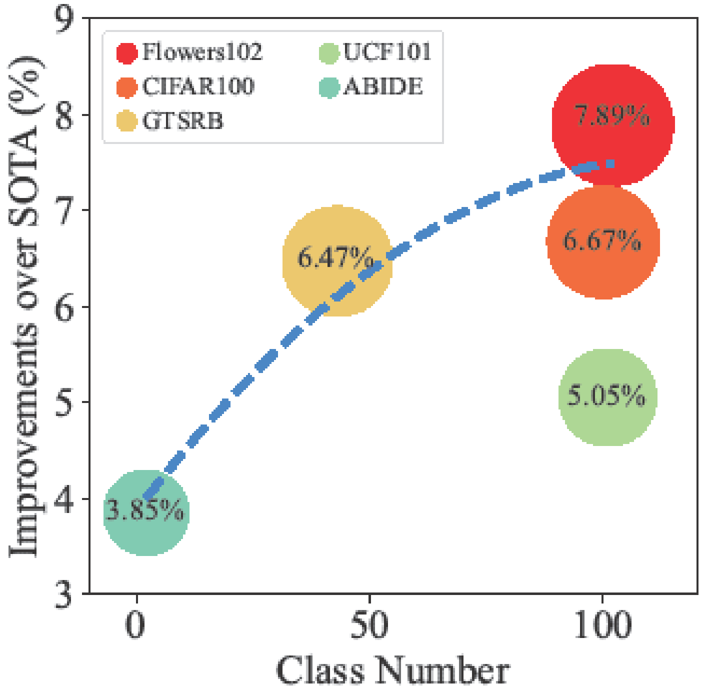


Figure 2. ILM-VP's improvements over FLM-VP on representative datasets (datasets with improvements over 3%) using ResNet-18.

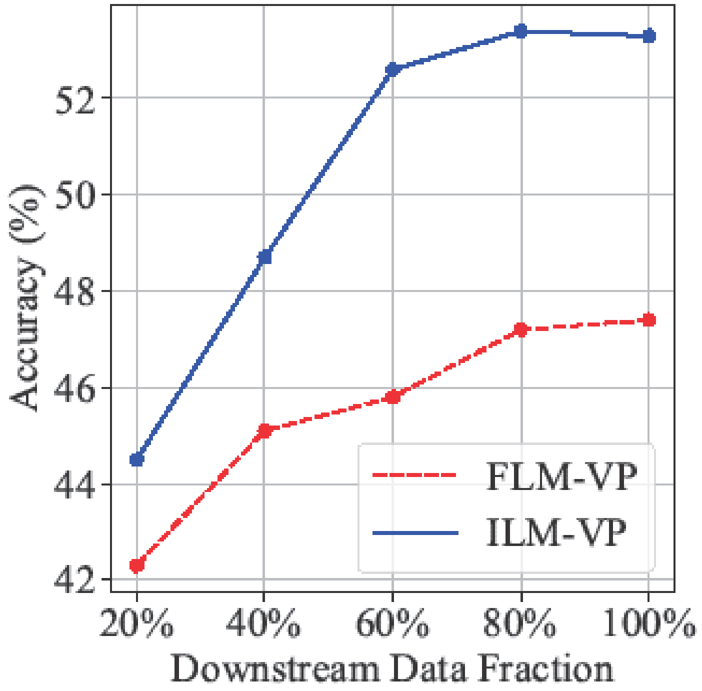
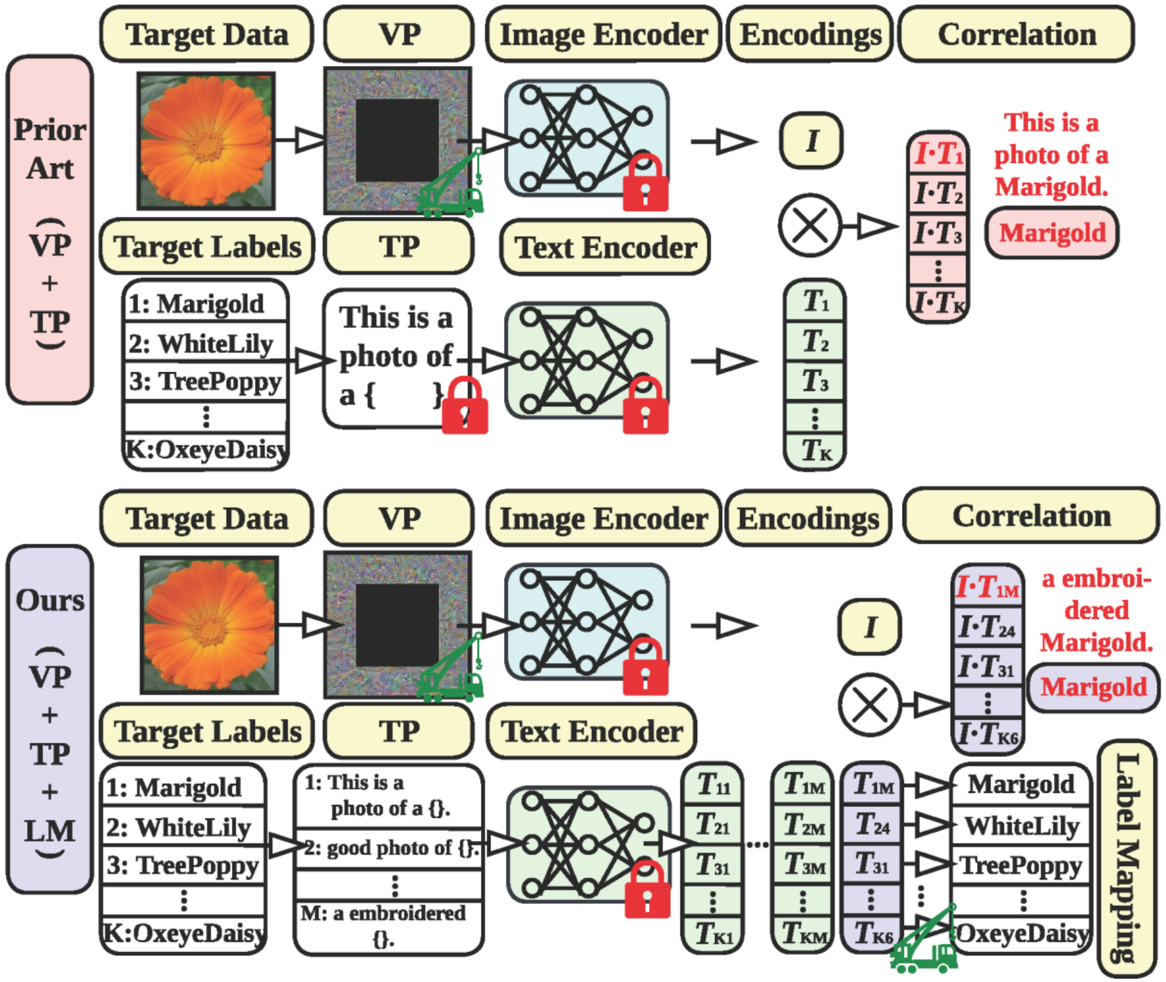


Figure 3. ILM-VP and FLM-VP performance on different fractions of GTSRB dataset (43 classes and more than 900 training samples per class) using ResNet-18.

Extension: LM in Text Domain for CLIP



Extension: LM in Text Domain for CLIP

Methods	VP+TP Acc(%)	Ours (VP+TP+LM) Acc(%)	Ours (VP+TP+LM) Examples of context prompt template → target label
Flowers102	70.0	**83.7**	a close-up photo of a {} → buttercup
DTD	56.8	**63.9**	graffiti of a {} → blotchy
UCF101	66.0	**70.6**	a {} in a video game → baseball pitch
Food101	78.9	**79.1**	a photo of the dirty {} → crab cake
SVHN	89.9	**91.2**	a photo of a {} → 7
EuroSAT	96.4	**96.9**	a pixelated photo of a {} → river
StanfordCars	57.2	**57.6**	the toy {} → 2011 audi s6 sedan
SUN397	60.5	**61.2**	a photo of a large {} → archive
CIFAR10	93.9	**94.4**	a pixelated photo of a {} → ship
ImageNet-R	67.5	**68.6**	a rendition of a {} → gold fish
ImageNet-Sketch	38.5	**39.7**	a sketch of a {} → eagle

Table 2. Results of CLIP-based prompt learning 'VP+TP+LM' and the baseline 'VP+TP' (restricted to using text prompt template "This is a photo of a {}") over 11 target datasets.

MICHIGAN STATE UNIVERSITY OPTML MIT-IBM Watson AI Lab JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

More Explanation Examples

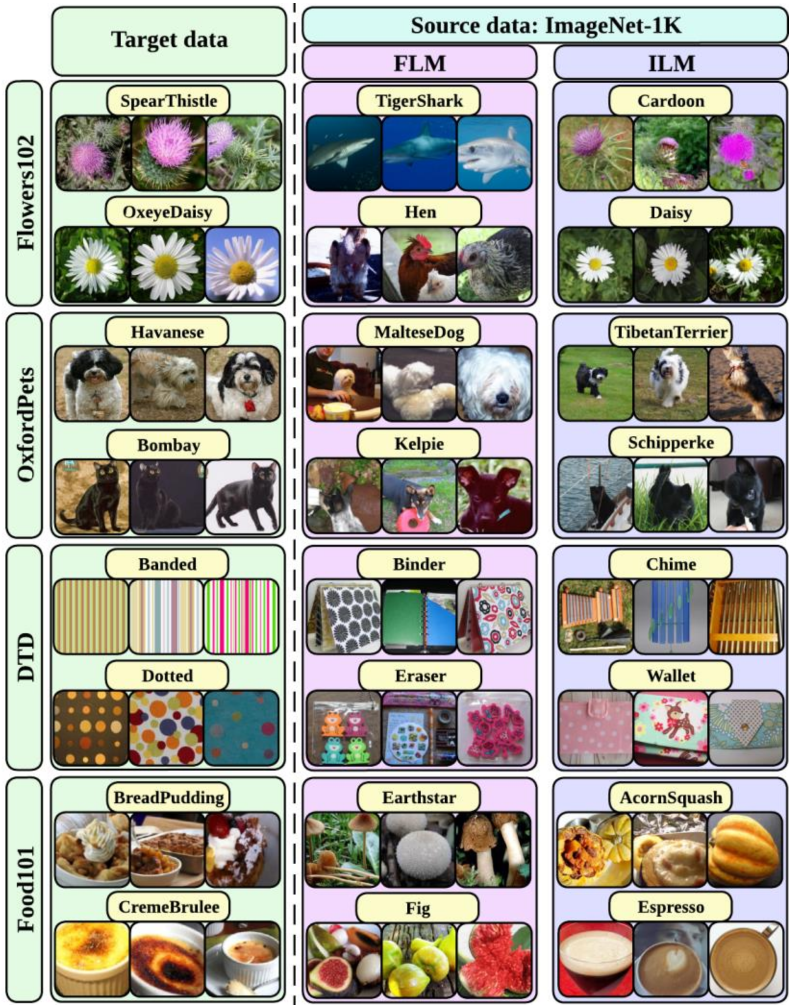


Figure 4. Interpretation merit of ILM vs. FLM, visualized by LM results in VP to re-purpose a ResNet-18 to conduct image classification tasks.

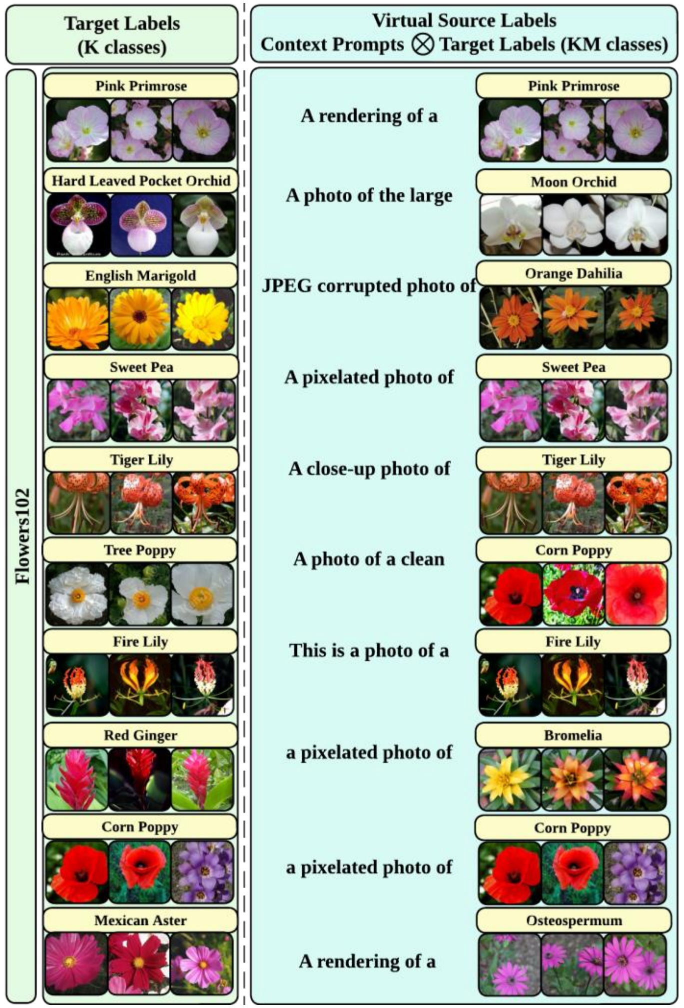


Figure 5. LM results of our proposed 'VP+TP+LM' method for CLIP, which shows significant interpretability.

MICHIGAN STATE UNIVERSITY
OPTML
MIT-IBM Watson AI Lab
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

terima kasih
multumesc
ありがとう
謝謝
baie
ngiyabonga
suksema
Met dank
obrigada
Thank You
dankie
molte grazie
감사합니다
Danke schön!
merci
obrigado
謝謝
gracias
Благодарность
شكرًا
Спасибі
Dziękuję
tusind tak
mahalo
dank u