CVPR 2022

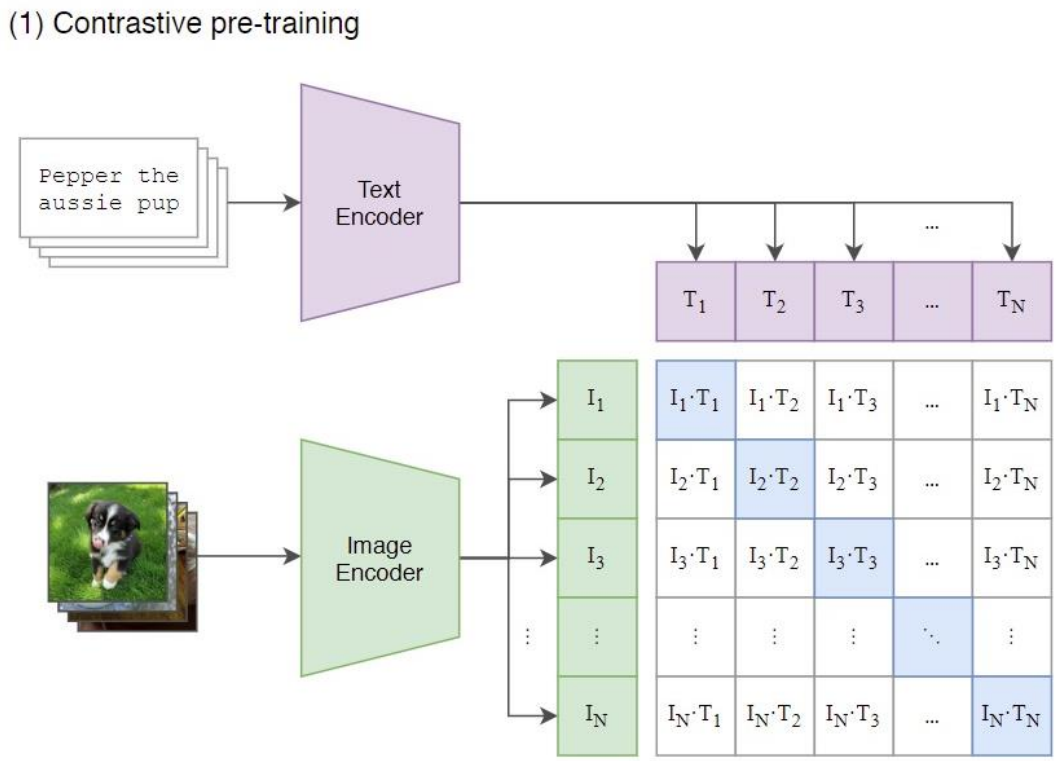
Conditional Prompt Learning for Vision-Language Models

Kaiyang Zhou Jingkang Yang Chen Change Loy Ziwei Liu ✉
S-Lab, Nanyang Technological University, Singapore
{kaiyang.zhou, jingkang001, ccloy, ziwei.liu}@ntu.edu.sg

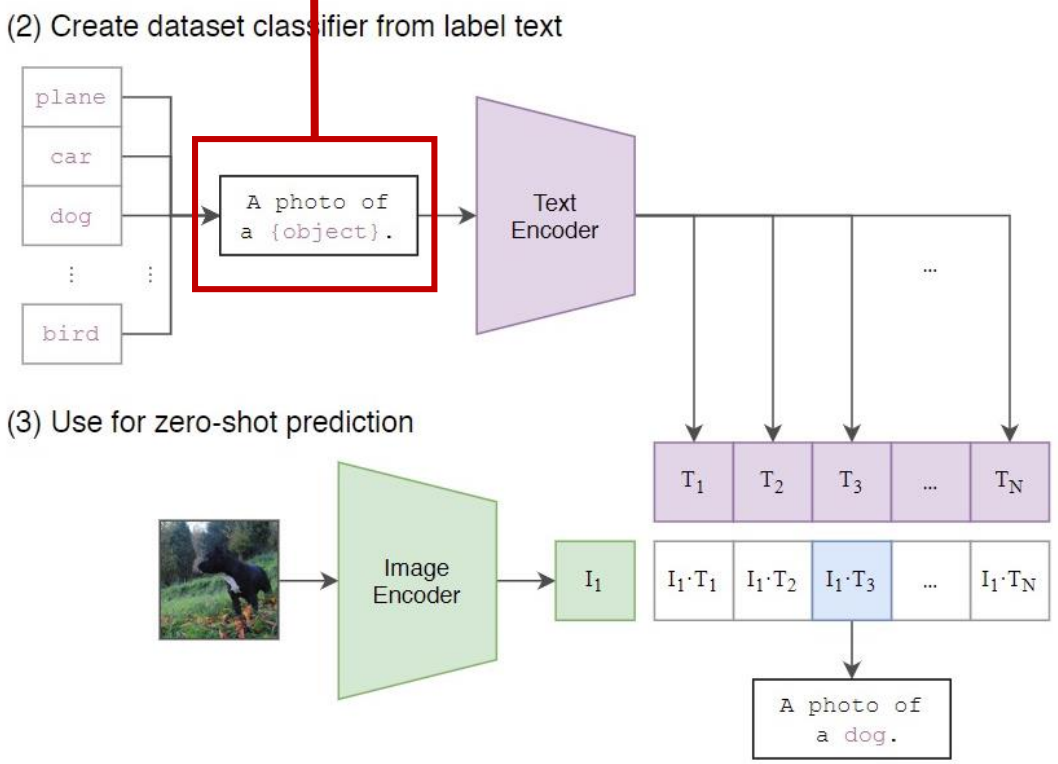
Mengxue

Pre—CLIP

- Prompt: 提示



Prompt: A photo of a _____.



Pre—CLIP

television studio (90.2%) Ranked 1 out of 397

- ✓ a photo of a **television studio**.
- × a photo of a **podium indoor**.
- × a photo of a **conference room**.
- × a photo of a **lecture room**.

A photo of a _____.

annual crop land (12.9%) Ranked 4 out of 10

- × a centered satellite photo of **permanent crop land**.
- × a centered satellite photo of **pasture land**.
- × a centered satellite photo of **highway or road**.
- ✓ a centered satellite photo of **annual crop land**.
- × a centered satellite photo of **brushland or shrubland**.

A center satellite photo of _____.

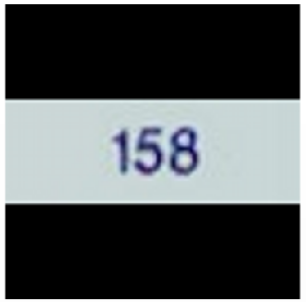
roundabout (96.4%) Ranked 1 out of 45

- ✓ satellite imagery of **roundabout**.
- × satellite imagery of **intersection**.
- × satellite imagery of **church**.
- × satellite imagery of **medium residential**.

Satellite imagery of _____.

STREET VIEW HOUSE NUMBERS (SVHN)

158 (0.3%) Ranked 83 out of 2000



- × a street sign of the number: "**1157**".
- × a street sign of the number: "**1165**".
- × a street sign of the number: "**1164**".
- × a street sign of the number: "**1155**".
- × a street sign of the number: "**1364**".

A street sign of the number:" _____" .

Pre—CoOp Context Optimization

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

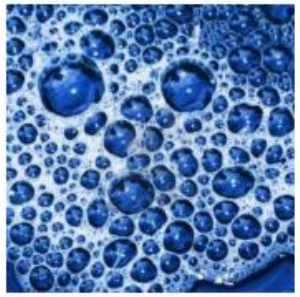
(a)

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)



Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

(c)

EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)

Fig. 1 Prompt engineering vs Context Optimization (CoOp). The former needs to use a held-out validation set for words tuning, which is inefficient; the latter automates the process and requires only a few labeled images for learning.

Learning to Prompt for Vision-Language Models. arXiv:2109.01134

Pre—CoOp Context Optimization

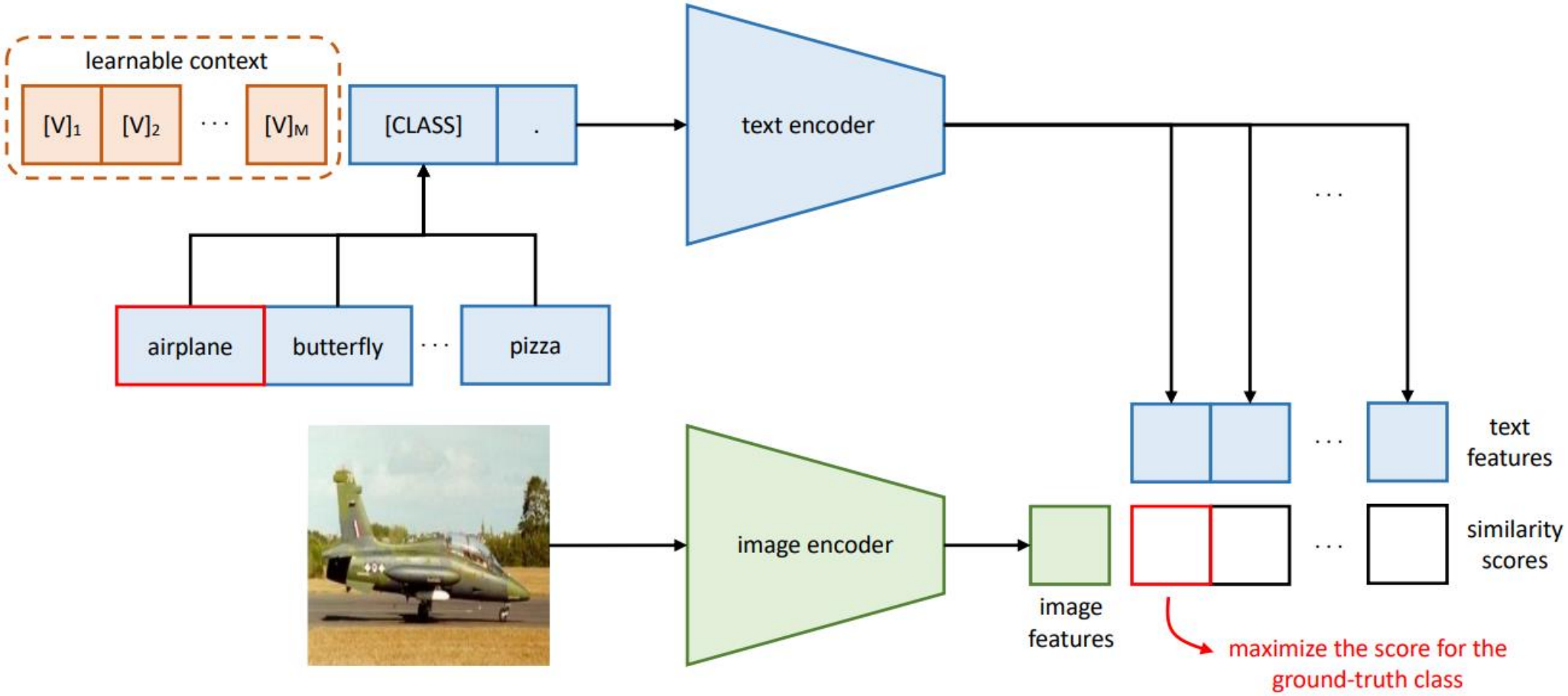
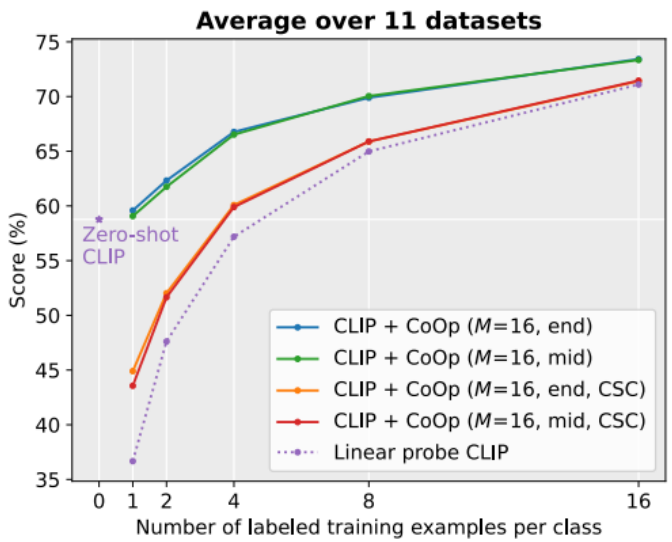
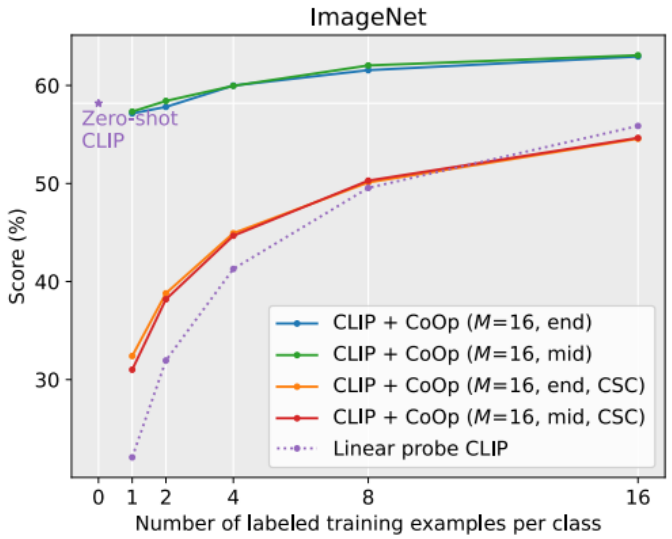
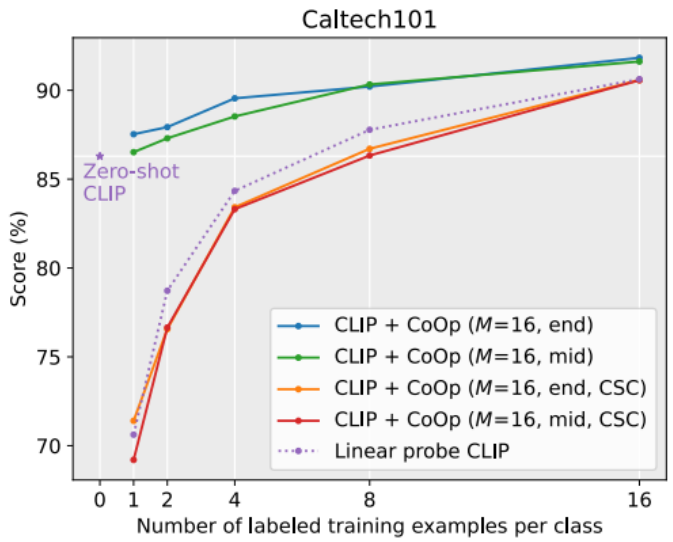


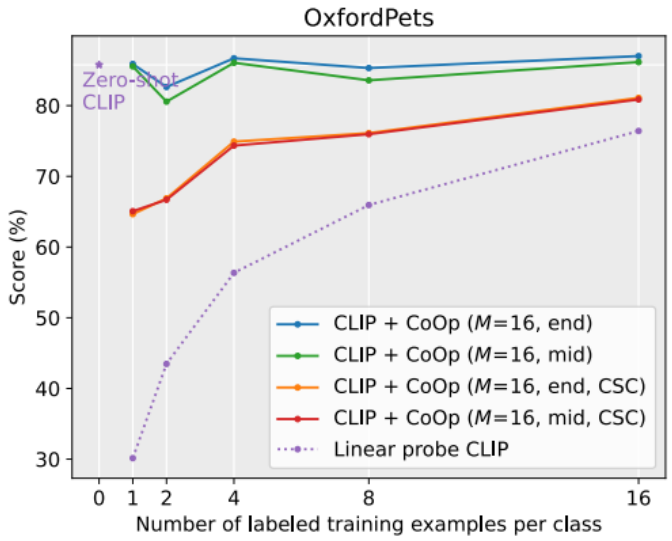
Fig. 2 Overview of Context Optimization (CoOp). The main idea is to model a prompt's context using a set of learnable vectors, which can be optimized through minimizing the classification loss. Two designs are proposed: one is unified context, which shares the same context vectors with all classes; and the other is class-specific context, which learns for each class a specific set of context vectors.

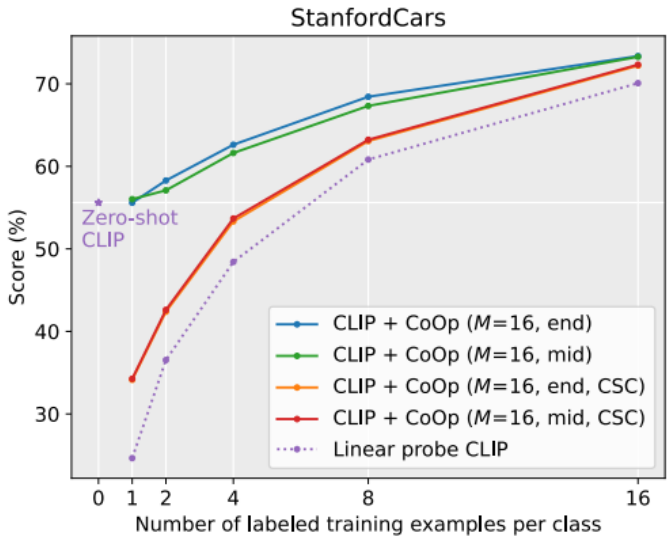
Pre—CoOp Context Optimization

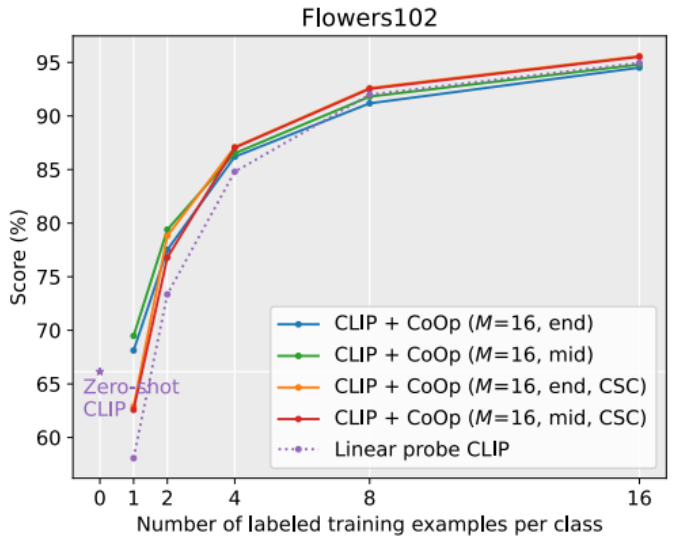




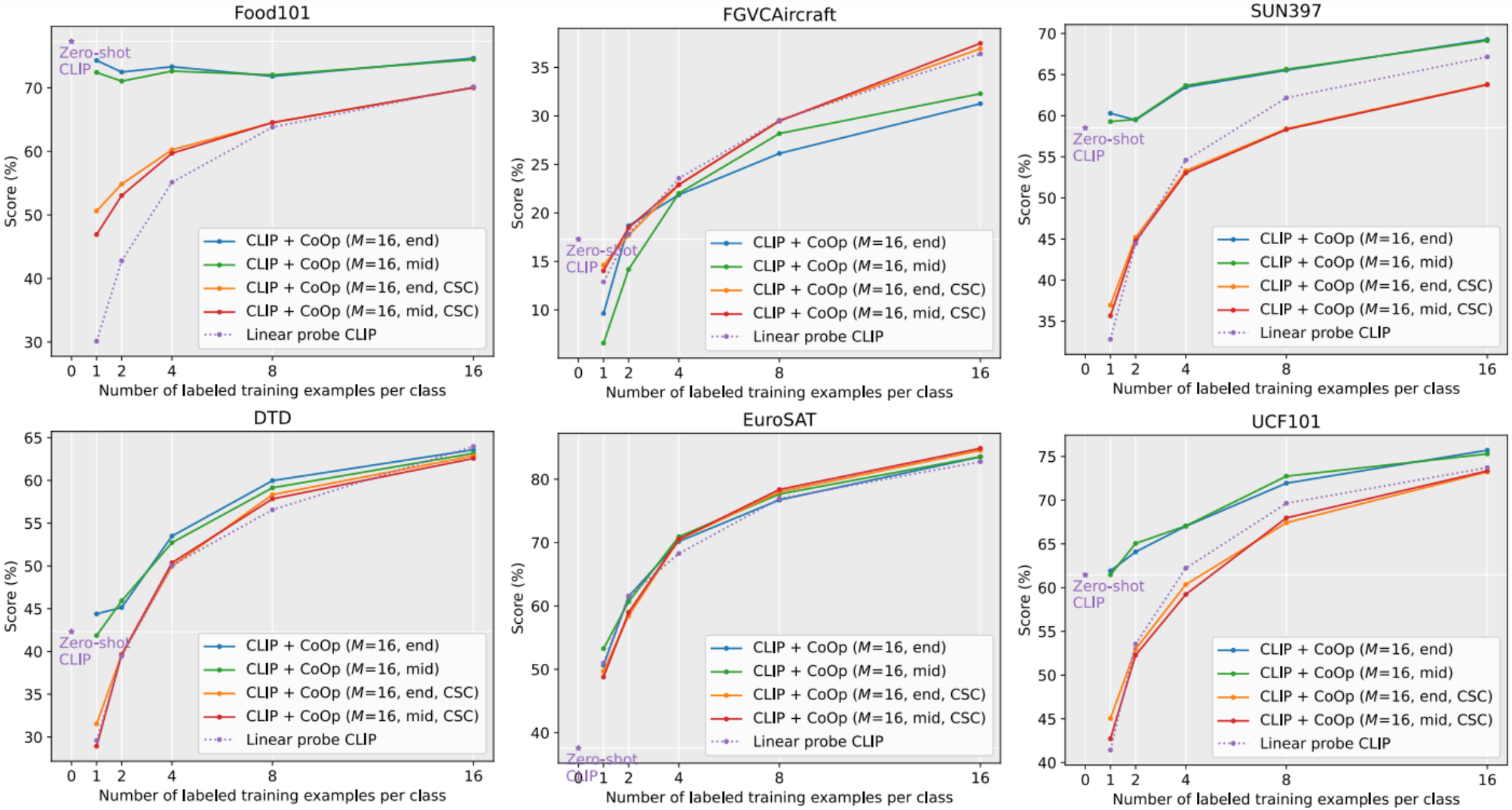








Pre—CoOp Context Optimization



Pre—CoOp Context Optimization

Table 1 Comparison with zero-shot CLIP on robustness to distribution shift using different vision backbones. M: CoOp's context length.

	Source	Target			
Method	ImageNet	-V2	-Sketch	-A	-R
ResNet-50					
Zero-Shot CLIP	58.18	51.34	33.32	21.65	56.00
Linear Probe CLIP	55.87	45.97	19.07	12.74	34.86
CLIP + CoOp ($M=16$)	62.95	55.11	32.74	22.12	54.96
CLIP + CoOp ($M=4$)	**63.33**	**55.40**	**34.67**	**23.06**	**56.60**
ResNet-101					
Zero-Shot CLIP	61.62	54.81	38.71	28.05	64.38
Linear Probe CLIP	59.75	50.05	26.80	19.44	47.19
CLIP + CoOp ($M=16$)	**66.60**	**58.66**	39.08	28.89	63.00
CLIP + CoOp ($M=4$)	65.98	58.60	**40.40**	**29.60**	**64.98**
ViT-B/32					
Zero-Shot CLIP	62.05	54.79	40.82	29.57	**65.99**
Linear Probe CLIP	59.58	49.73	28.06	19.67	47.20
CLIP + CoOp ($M=16$)	**66.85**	58.08	40.44	30.62	64.45
CLIP + CoOp ($M=4$)	66.34	**58.24**	**41.48**	**31.34**	65.78
ViT-B/16					
Zero-Shot CLIP	66.73	60.83	46.15	47.77	73.96
Linear Probe CLIP	65.85	56.26	34.77	35.68	58.43
CLIP + CoOp ($M=16$)	**71.92**	64.18	46.71	48.41	74.32
CLIP + CoOp ($M=4$)	71.73	**64.56**	**47.89**	**49.93**	**75.14**

Pre—CoOp Context Optimization

#	ImageNet	Food101	OxfordPets	DTD
1	potd (1.7136)	lc (0.6752)	tosc (2.5952)	boxed (0.9433)
2	that (1.4015)	enjoyed (0.5305)	judge (1.2635)	seed (1.0498)
3	filmed (1.2275)	beh (0.5390)	fluffy (1.6099)	anna (0.8127)
4	fruit (1.4864)	matches (0.5646)	cart (1.3958)	mountain (0.9509)
5	,... (1.5863)	nytimes (0.6993)	harlan (2.2948)	eldest (0.7111)
6	° (1.7502)	prou (0.5905)	paw (1.3055)	pretty (0.8762)
7	excluded (1.2355)	lower (0.5390)	incase (1.2215)	faces (0.7872)
8	cold (1.4654)	N/A	bie (1.5454)	honey (1.8414)
9	stery (1.6085)	minute (0.5672)	snuggle (1.1578)	series (1.6680)
10	warri (1.3055)	~ (0.5529)	along (1.8298)	coca (1.5571)
11	marvelcomics (1.5638)	well (0.5659)	enjoyment (2.3495)	moon (1.2775)
12	.: (1.7387)	ends (0.6113)	jt (1.3726)	lh (1.0382)
13	N/A	mis (0.5826)	improving (1.3198)	won (0.9314)
14	lation (1.5015)	somethin (0.6041)	srsly (1.6759)	replied (1.1429)
15	muh (1.4985)	seminar (0.5274)	asteroid (1.3395)	sent (1.3173)
16	.# (1.9340)	N/A	N/A	piedmont (1.5198)

Table 3: Random vs. manual initialization.

	Avg %
$[V]_1[V]_2[V]_3[V]_4$	71.26
"a photo of a"	**71.51**

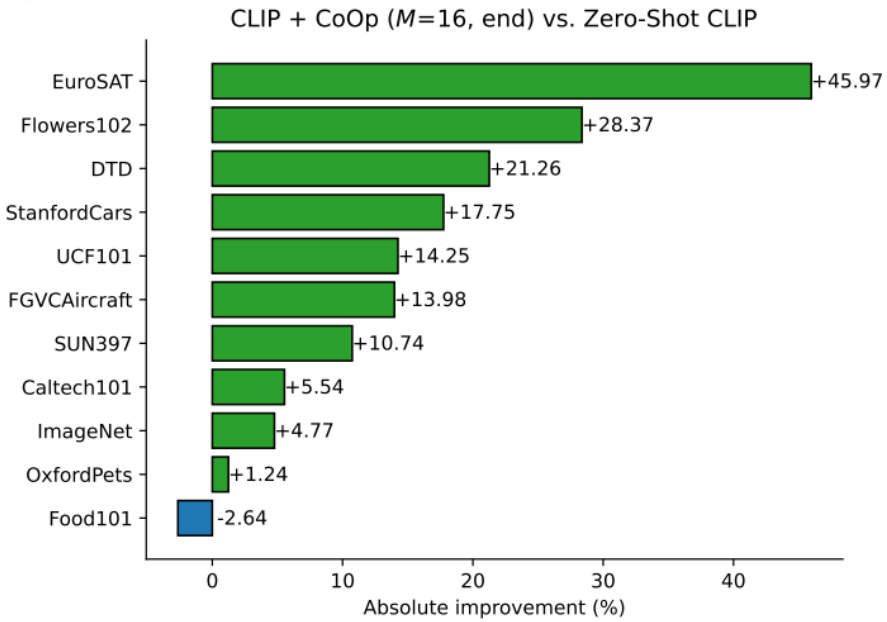
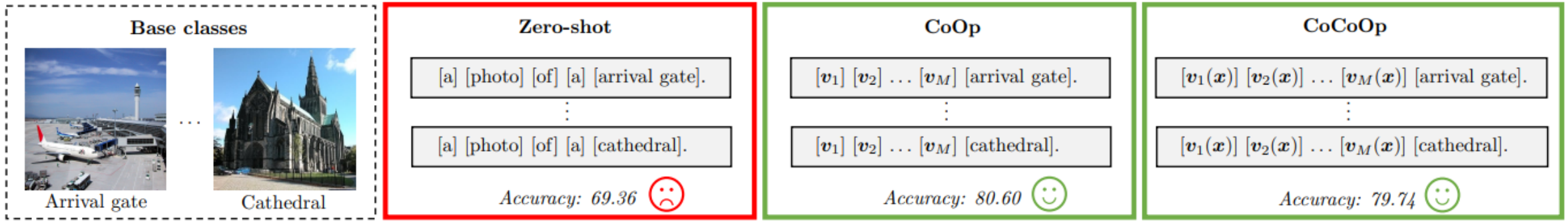


Fig. 4 Comparison with hand-crafted prompts.

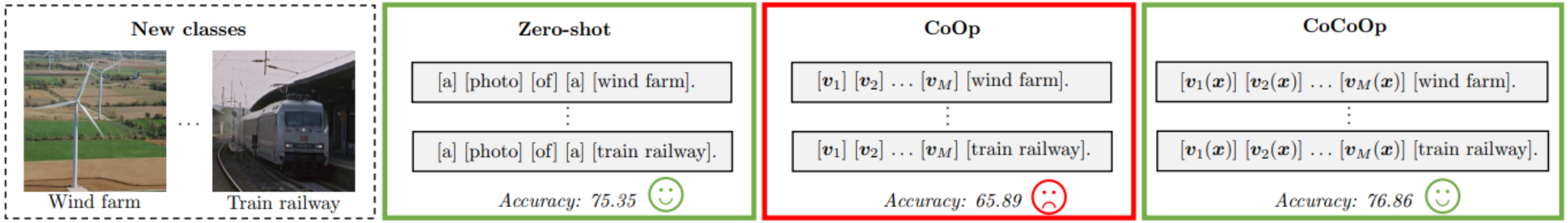
Motivation—CoCoOp

Conditional Context Optimization

- To learn generalizable prompts.



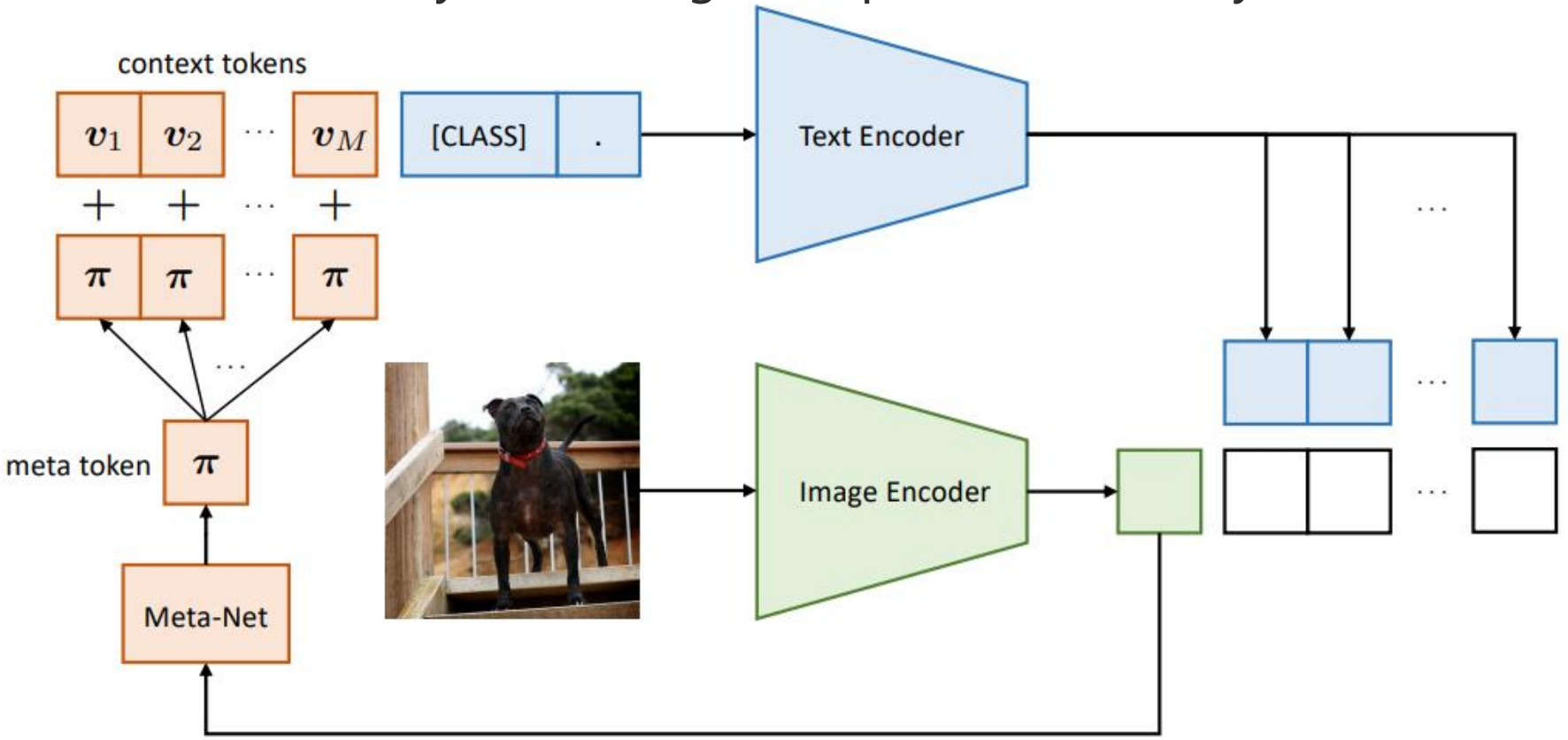
(a) Both CoOp and CoCoOp work well on the base classes observed during training and beat manual prompts by a significant margin.



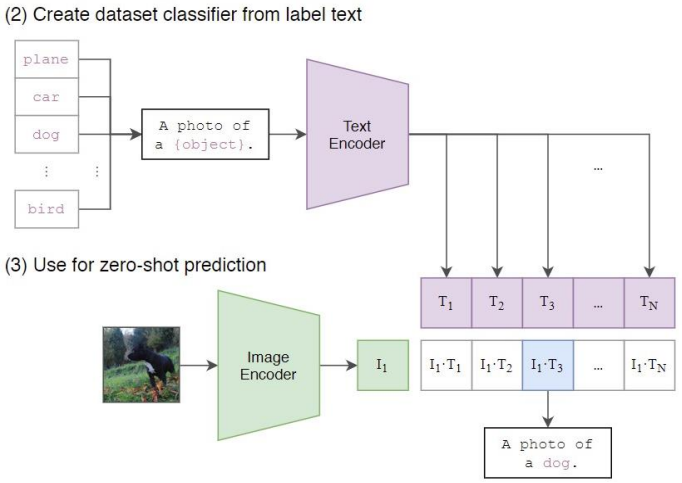
(b) The instance-conditional prompts learned by CoCoOp are much more generalizable than CoOp to the unseen classes.

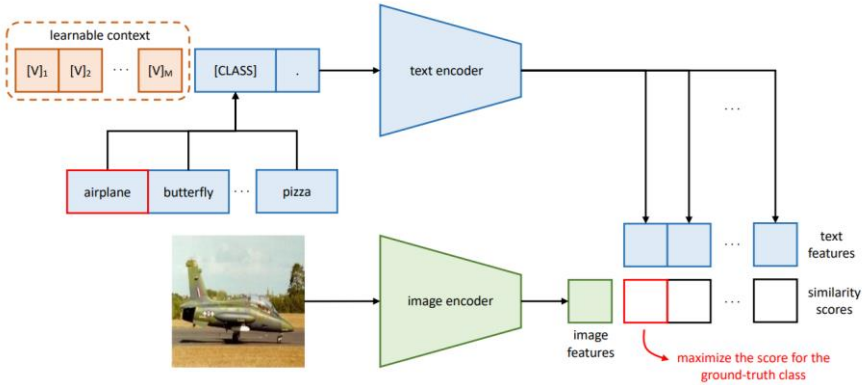
Framework—CoCoOp

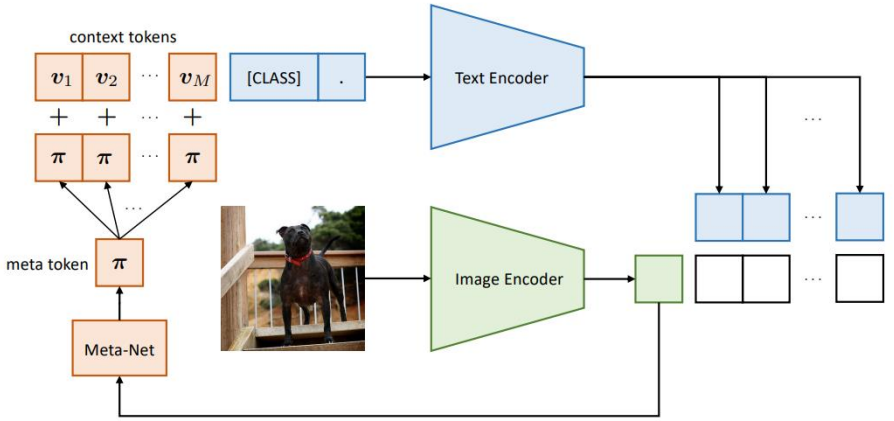
- In this work, the Meta-Net is built with a two-layer bottleneck structure (Linear-ReLU-Linear), with the hidden layer reducing the input dimension by 16×.



Comparison of CLIP/CoOp/CoCoOp







$$p(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{x}, \boldsymbol{w}_y)/\tau)}{\sum_{i=1}^{K} \exp(\text{sim}(\boldsymbol{x}, \boldsymbol{w}_i)/\tau)},$$

$$p(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_y))/\tau)}{\sum_{i=1}^{K} \exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_i)/\tau)}.$$

$$p(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_y(\boldsymbol{x})))/\tau)}{\sum_{i=1}^{K} \exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_i(\boldsymbol{x}))/\tau)}.$$

Generalization From Base to New Classes

- Split the classes equally into two groups: base classes and new classes
- Trained on base classes
- Evaluated on the base and new classes separately to test generalizability
- 16 shots
- Vision backbone in CLIP, i.e., ViT-B/16
- Fix the context length to 4 and initialize the context vectors using the pre-trained word embeddings of "a photo of a" for both CoOp and CoCoOp.

Generalization From Base to New Classes

(a) **Average over 11 datasets.**

	Base	New	H
CLIP	69.34	**74.22**	71.70
CoOp	**82.69**	63.22	71.66
CoCoOp	80.47	71.69	**75.83**

(b) ImageNet.

	Base	New	H
CLIP	72.43	68.14	70.22
CoOp	**76.47**	67.88	71.92
CoCoOp	75.98	**70.43**	**73.10**

(c) Caltech101.

	Base	New	H
CLIP	96.84	**94.00**	95.40
CoOp	**98.00**	89.81	93.73
CoCoOp	97.96	93.81	**95.84**

(d) OxfordPets.

	Base	New	H
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	**95.20**	**97.69**	**96.43**

(e) StanfordCars.

	Base	New	H
CLIP	63.37	**74.89**	68.65
CoOp	**78.12**	60.40	68.13
CoCoOp	70.49	73.59	**72.01**

(f) Flowers102.

	Base	New	H
CLIP	72.08	**77.80**	74.83
CoOp	**97.60**	59.67	74.06
CoCoOp	94.87	71.75	**81.71**

(g) Food101.

	Base	New	H
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	**90.70**	**91.29**	**90.99**

(h) FGVCAircraft.

	Base	New	H
CLIP	27.19	**36.29**	**31.09**
CoOp	**40.44**	22.30	28.75
CoCoOp	33.41	23.71	27.74

(i) SUN397.

	Base	New	H
CLIP	69.36	75.35	72.23
CoOp	**80.60**	65.89	72.51
CoCoOp	79.74	**76.86**	**78.27**

(j) DTD.

	Base	New	H
CLIP	53.24	**59.90**	56.37
CoOp	**79.44**	41.18	54.24
CoCoOp	77.01	56.00	**64.85**

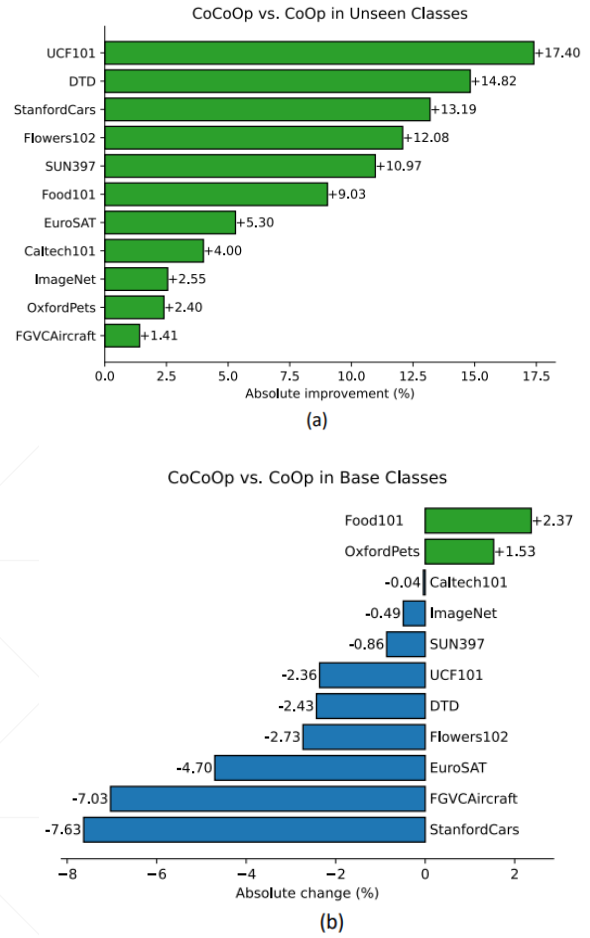
(k) EuroSAT.

	Base	New	H
CLIP	56.48	**64.05**	60.03
CoOp	**92.19**	54.74	68.69
CoCoOp	87.49	60.04	**71.21**

(l) UCF101.

	Base	New	H
CLIP	70.53	**77.50**	73.85
CoOp	**84.69**	56.05	67.46
CoCoOp	82.33	73.45	**77.64**

H: Harmonic mean (to highlight the generalization trade-off [54]).



$$H = 2 * (acc_{\mathcal{Y}^{tr}} * acc_{\mathcal{Y}^{ts}})/(acc_{\mathcal{Y}^{tr}} + acc_{\mathcal{Y}^{ts}}) \quad (16)$$

where $acc_{\mathcal{Y}^{tr}}$ and $acc_{\mathcal{Y}^{ts}}$ represent the accuracy of images from seen ($\mathcal{Y}^{tr}$), and images from unseen ($\mathcal{Y}^{ts}$) classes re-

Cross-Dataset Transfer

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp [62]	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	**46.39**	66.55	63.88
CoCoOp	71.02	**94.43**	**90.14**	**65.32**	**71.88**	**86.06**	**22.94**	**67.36**	**45.73**	45.37	**68.21**	**65.74**
Δ	**-0.49**	**+0.73**	**+1.00**	**+0.81**	**+3.17**	**+0.76**	**+4.47**	**+3.21**	**+3.81**	**-1.02**	**+1.66**	**+1.86**

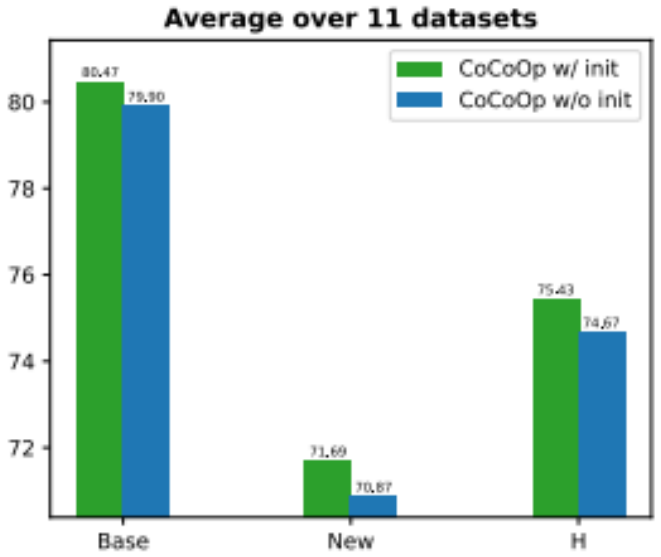
Domain Generalization

		Source	Target			
	Learnable?	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CLIP [40]		66.73	60.83	46.15	47.77	73.96
CoOp [62]	✓	**71.51**	**64.20**	47.99	49.71	75.21
CoCoOp	✓	71.02	64.07	**48.75**	**50.63**	**76.18**

Further Analysis

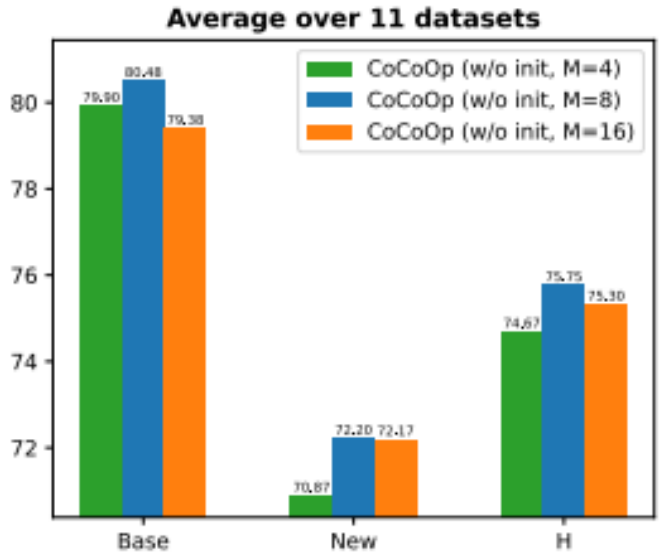
Class-Incremental Test

	Learnable?	Accuracy
CLIP [40]		65.22
CoOp [62]	✓	65.55
CoCoOp	✓	**69.13**



(a) Ablation on initialization.

Model	# params	Base	New	H
CoOp (ctx=4)	2,048	**76.47**	67.88	71.92
CoOp (ctx=60)	30,720	76.16	65.34	70.34
CoOp (ctx=4) + Meta-Net	34,816	75.98	**70.43**	**73.10**



(b) Ablation on context length.

Limitations

- **The first limitation is about training efficiency**: CoCoOp is slow to train and would consume a significant amount of GPU memory if the batch size is set larger than one. The reason is because CoCoOp is based on an instanceconditional design that requires for each image an independent forward pass of instance-specific prompts through the text encoder. This is much less efficient than CoOp that only needs a single forward pass of prompts through the text encoder for an entire mini-batch of any size.

- **The second limitation is that on 7 out of the 11 datasets (see Table 1), CoCoOp' s performance in unseen classes still lags behind CLIP' s**, indicating that more efforts are needed from the community to fully close or overturn the gaps between manual and learning-based prompts.

Thanks