Read-only Prompt Optimization for Vision-Language Few-shot Learning (ICCV 2023)

Dongjun Lee∗ Seokwon Song∗ Jihee Suh
Joonmyeong Choi Sanghyeok Lee Hyunwoo J. Kim†
Korea University

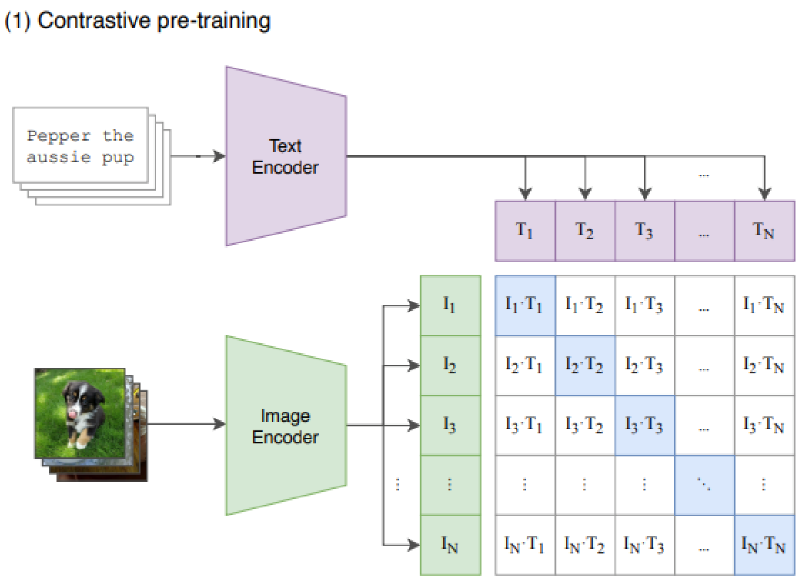
Yurim Lee

yurimmy65@gmail.com

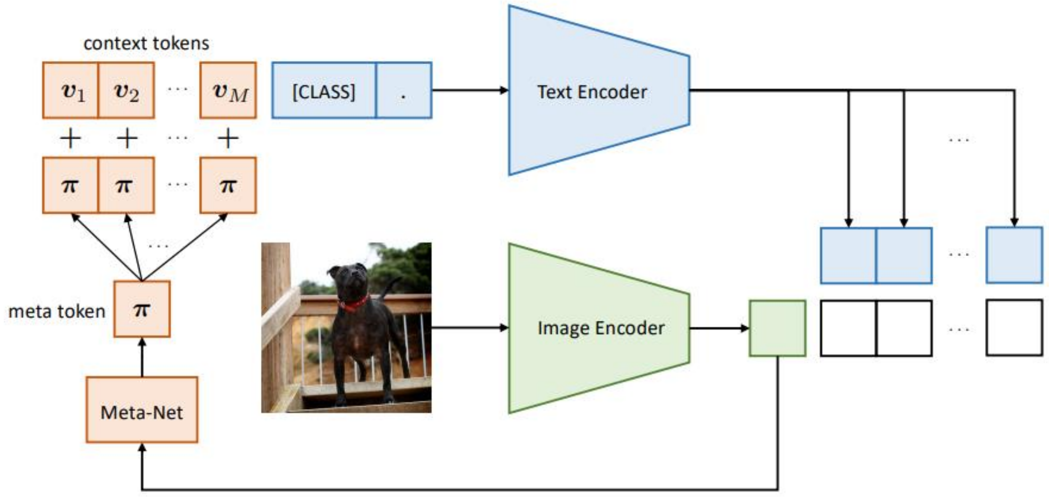
2024. 2. 7

고려대학교
KOREA UNIVERSITY

Ⅰ. Introduction

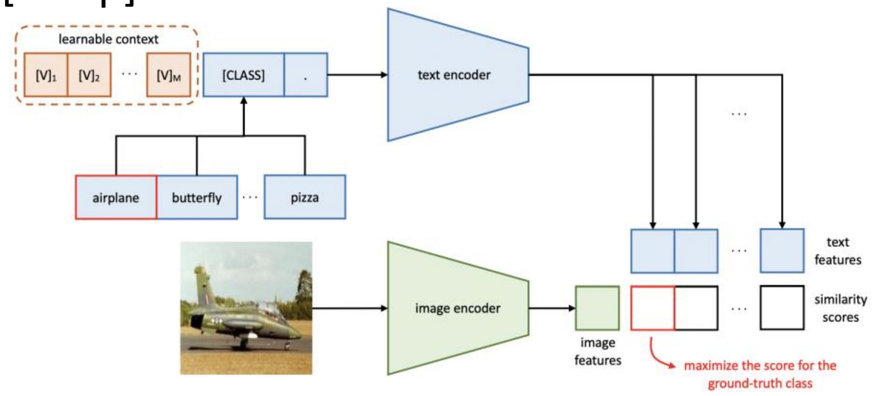


[CLIP]

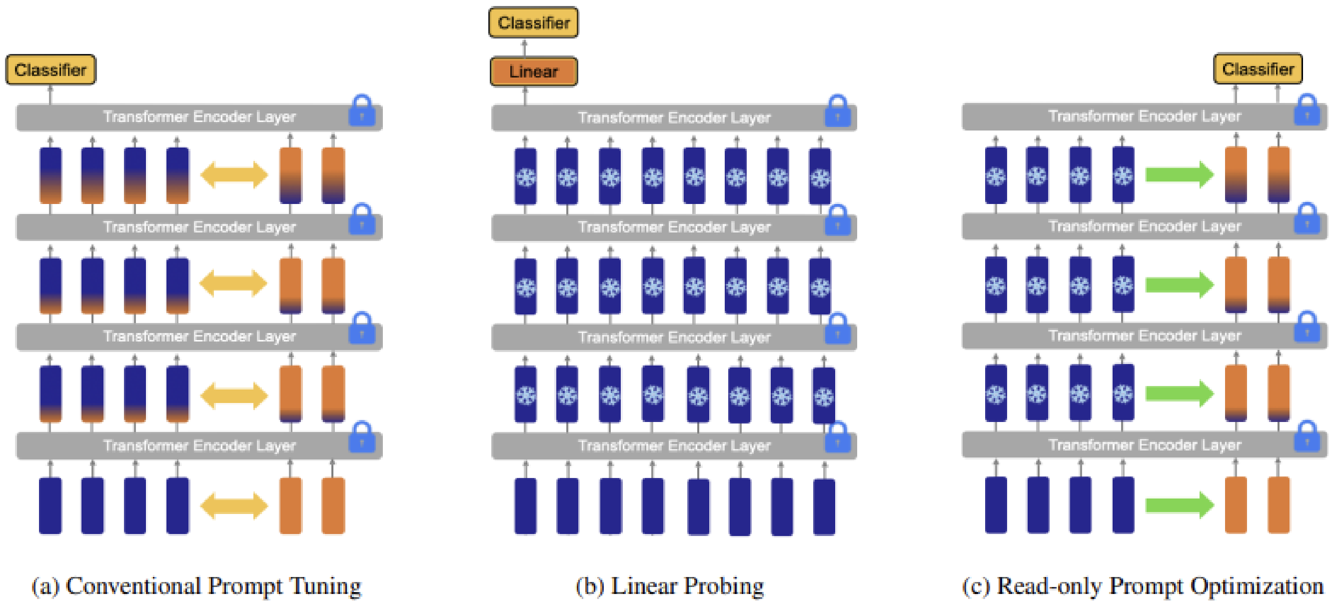


[CoCoOp]

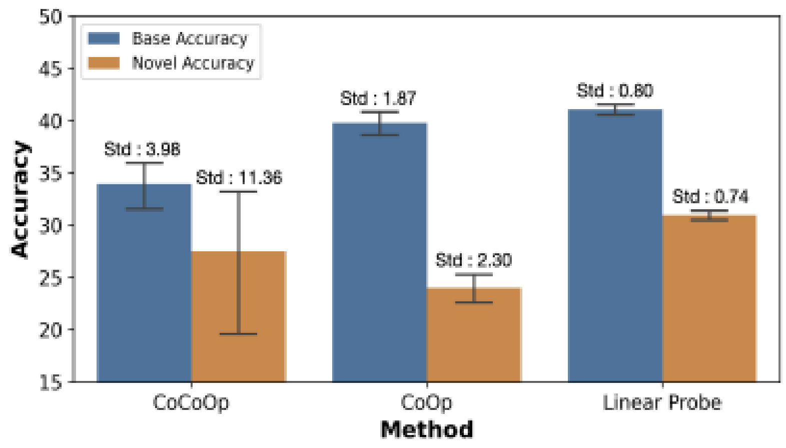
[CoOp]



Ⅰ. Introduction



"Internal Representation Shift"



CoOp, CoCoOp: High Variance
-> may have negatively impact robustness & generalization in data-deficient setting

Linear probing: Parameter inefficient(262k), Lack of generalizability in domain-shift task

Ⅰ. Introduction

Contributions

• We propose Read-only Prompt Optimization (RPO), which allows prompts only to read information from the attention-based interactions of a pre-trained vision language model, thereby preventing the internal representation shift.

• We develop a simple yet effective initialization method for our read-only prompts, leveraging the special token embeddings of the pre-trained CLIP vision-language model.

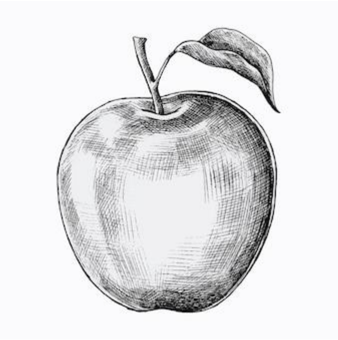
• Our extensive experiments and analyses demonstrate the generalization of RPO on domain and label shift in few-shot adaptation settings, achieving the best performance in 9 benchmarks on base to new generalization and in 4 benchmarks on domain generalization, at the same time reducing variance depending on the few-shot sample

Ⅱ. Related works

Vision-Language Models
Especially good on zero-shot image classification
But, adapting to specific tasks is challenging

Prompt Learning
Incorporating additional tokens (handcrafted instruction, learnable prompts)
Visual prompt & Text prompt

Zero-shot Learning & Domain Generalization
Learning general knowledge from ‘base’ -> adapt to novel classes
Domain – invariant representations are needed



Ⅲ. Method

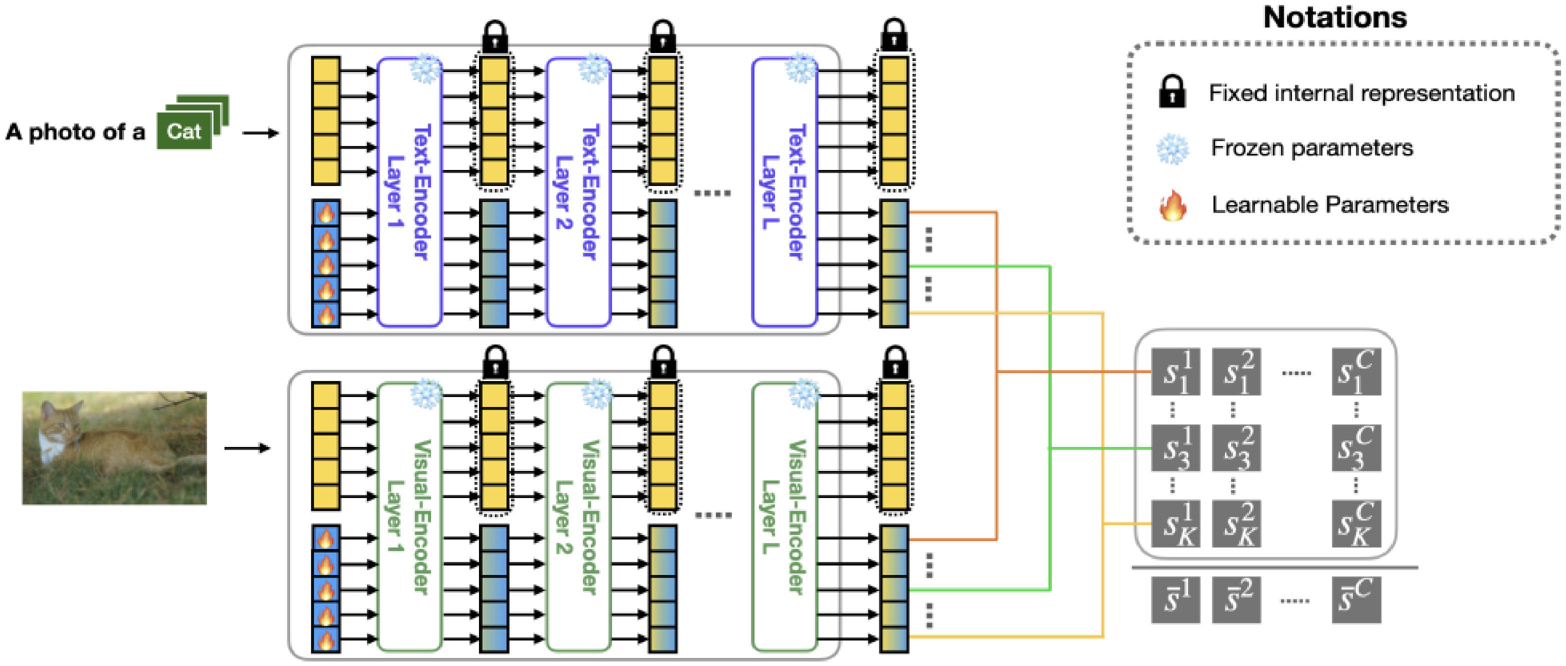


Figure 3: **Overall architecture of RPO.** We use the default prompt "A photo of a [CLASS]" for all datasets. Then in both encoders, our read-only prompts are concatenated to the original features and fed into a frozen encoder. Attention within these encoders are masked so that our prompts can be learned, but not shift the original feature interactions. We compute similarity scores between the outputs of each encoder corresponding to each of K prompts and average them to produce final classification scores $\bar{s}^1$ to $\bar{s}^C$, where C denotes the number of classes.

Ⅲ. Method

3.1. Read-only Prompts

$$\mathbf{x}^{(0)} = \left[x^{(0)}; E_x^{(0)}; \{p_i^v\}_{i=1}^K\right], \qquad (1)$$

<Visual prompt>
[Special Token embedding(CLS) ; Visual embedding; ith learnable prompt] , K= number of Prompts

$$\mathbf{y}^{(0)} = \left[y^{(0)}; E_y^{(0)}; \{p_i^t\}_{i=1}^K\right], \qquad (2)$$

<Text prompt>
[Special Token embedding(EOS) ; Text embedding; ith learnable prompt]

Ⅲ. Method

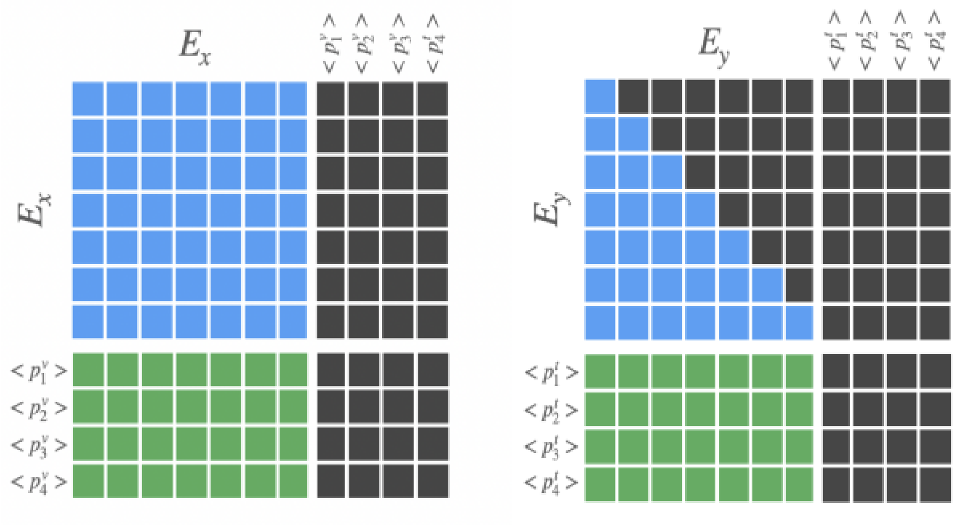
3.2. Special token-based initialization

$$p_i^v \sim \mathcal{N}(x^{(0)}, \sigma^2 I), \quad p_i^t \sim \mathcal{N}(y^{(0)}, \sigma^2 I), \qquad (3)$$

Visual Encoder: [CLS]
Text Encoder: [EOS]
: Feature aggregator
σ = 0.1 -> Avoid constant initialization

Ⅲ. Method

3.3. Masked attention



(a) Visual Attention Mask (b) Textual Attention Mask

Figure 4: The visualization of attention masks for each encoder.

[Mask]

$$M_v^{i,j} = \begin{cases} -\infty, & \text{if } j > 1 + N_x \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

$$M_t^{i,j} = \begin{cases} -\infty, & \text{if } j > 1 + N_y \text{ or } i > j \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

[Masked attention Operation]

$$\begin{aligned} \mathbf{x}^{(l+1)} &= \mathcal{V}_{l+1}(\mathbf{x}^{(l)}, M_v) \\ &= \mathbf{softmax}\left(\frac{QK^T}{\sqrt{d_v}} + M_v\right) \cdot V, \\ \mathbf{y}^{(l+1)} &= \mathcal{T}_{l+1}(\mathbf{y}^{(l)}, M_t) \\ &= \mathbf{softmax}\left(\frac{QK^T}{\sqrt{d_t}} + M_t\right) \cdot V, \end{aligned} \quad (6)$$

[Final outputs]

$$\begin{aligned} \mathbf{x}^{(L)} &= \left[e_0; E_x^{(L)}; \{e_i\}_{i=1}^{K}\right], \\ \mathbf{y}^{(L)} &= \left[s_0; E_y^{(L)}; \{s_i\}_{i=1}^{K}\right], \end{aligned} \quad (7)$$

$$\begin{aligned} v_i &= \mathbf{P}_v \cdot e_i, \\ t_i &= \mathbf{P}_t \cdot s_i, \end{aligned} \quad (8)$$

Ⅲ. Method

3.4. Pairwise Scoring Function

$$\mathrm{sim}(x, y) = \frac{1}{K} \sum_{i=1}^{K} \frac{v_i \cdot t_i}{|v_i||t_i|} \tag{9}$$

$$p(y_k|x) = \frac{\exp(\mathrm{sim}(x, y_k)/\tau)}{\sum_{j=1}^{C} \exp(\mathrm{sim}(x, y_j)/\tau)} \tag{10}$$

Averaging logits -> same effect as an ensemble of K independent models (separate perspectives about image & text)

Ⅳ. Experiments

(a) Average over 11 datasets

Methods	Base	Novel	H
CLIP	69.34	74.22	71.70
+LP	81.80	69.17	74.65
+CoOp	82.69	63.22	71.66
+CoCoOp	80.47	71.69	75.83
+RPO	81.13	75.00	**77.78**

(b) ImageNet.

Methods	Base	Novel	H
CLIP	72.43	68.14	70.22
+LP	73.13	57.10	64.13
+CoOp	76.47	67.88	71.92
+CoCoOp	75.98	70.43	73.10
+RPO	76.60	71.57	**74.00**

(c) Caltech101.

Methods	Base	Novel	H
CLIP	96.84	94.00	95.40
+LP	98.03	93.50	95.71
+CoOp	98.00	89.81	93.73
+CoCoOp	97.96	93.81	95.84
+RPO	97.97	94.37	**96.03**

(d) OxfordPets.

Methods	Base	Novel	H
CLIP	91.17	97.26	94.12
+LP	94.87	92.50	93.67
+CoOp	93.67	95.29	94.47
+CoCoOp	95.20	97.69	**96.43**
+RPO	94.63	97.50	96.05

(e) StanfordCars.

Methods	Base	Novel	H
CLIP	63.37	74.89	68.65
+LP	78.60	65.50	71.45
+CoOp	78.12	60.40	68.13
+CoCoOp	70.49	73.59	72.01
+RPO	73.87	75.53	**74.69**

(f) Flowers102.

Methods	Base	Novel	H
CLIP	72.08	77.08	74.83
+LP	97.87	65.87	78.74
+CoOp	97.60	59.67	74.06
+CoCoOp	94.87	71.75	81.71
+RPO	94.13	76.67	**84.50**

(g) Food101.

Methods	Base	Novel	H
CLIP	90.10	91.22	90.66
+LP	88.30	88.03	88.17
+CoOp	88.33	82.26	85.19
+CoCoOp	90.70	91.29	**90.99**
+RPO	90.33	90.83	90.58

(h) FGVCAircraft.

Methods	Base	Novel	H
CLIP	27.19	36.29	31.09
+LP	41.37	31.13	35.53
+CoOp	40.44	22.30	28.75
+CoCoOp	33.41	23.71	27.74
+RPO	37.33	34.20	**35.70**

(i) SUN397.

Methods	Base	Novel	H
CLIP	69.36	75.35	72.23
+LP	79.47	69.73	74.28
+CoOp	80.60	65.89	72.51
+CoCoOp	79.74	76.86	78.27
+RPO	80.60	77.80	**79.18**

(j) DTD.

Methods	Base	Novel	H
CLIP	53.24	59.90	56.37
+LP	80.63	55.97	66.07
+CoOp	79.44	41.18	54.24
+CoCoOp	77.01	56.00	64.85
+RPO	76.70	62.13	**68.61**

(k) EuroSAT.

Methods	Base	Novel	H
CLIP	56.48	64.05	60.03
+LP	82.30	68.00	74.47
+CoOp	92.19	54.74	68.69
+CoCoOp	87.49	60.04	71.21
+RPO	86.63	68.97	**76.79**

(l) UCF101.

Methods	Base	Novel	H
CLIP	70.53	77.50	73.85
+LP	85.27	73.53	78.97
+CoOp	84.69	56.05	67.46
+CoCoOp	82.33	73.45	77.64
+RPO	83.67	75.43	**79.34**

Ⅳ. Experiments

Table 2: **Comparison of RPO, CoCoOp, CoOp and manual prompt in domain generalization.** RPO learns from ImageNet (16 images per class) and is evaluated by 4 datasets with distribution shift and ImageNet itself. RPO performs better on 4 out of 5 datasets compared to CoCoOp.

		Source	Target			
	Learnable?	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CLIP		66.73	60.83	46.15	47.77	73.96
+CoOp	✓	71.51	64.20	47.99	49.71	75.21
+CoCoOp	✓	71.02	64.07	48.75	**50.63**	76.18
+RPO	✓	**71.67**	**65.13**	**49.27**	50.13	**76.57**

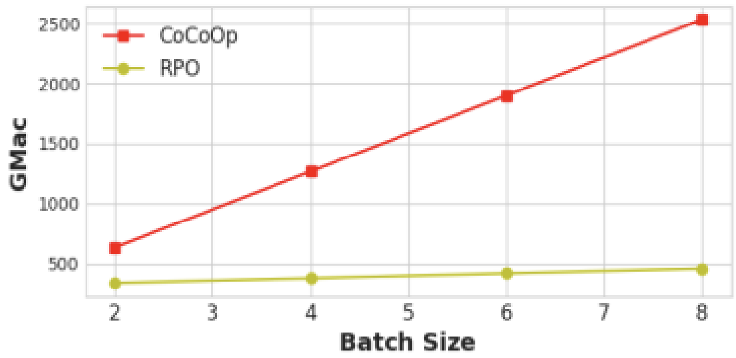


Figure 6: **Computational Cost of CoCoOp an RPO.**

IV. Experiments

4.3. Analysis

[Masked attention & ST-initialization]

Table 3: **Ablation result averaged over 11 datasets.**

Methods	Base	Novel	H
RPO w.o mask/init	78.63	69.56	73.29
RPO w.o mask	78.55	71.34	74.59
RPO w.o init	82.00	72.94	76.82
RPO	81.13	75.00	**77.78**

Table 4: **Analysis of RPO on extreme few shot settings.** We report RPO's averaged base accuracy, novel accuracy, and their harmonic mean on 10 benchmark datasets. RPO consistently outperforms CoCoOp on 1, 2, 4, and 8 shot setting evaluated by harmonic mean.

	1 shot		2 shot		4 shot		8 shot		16 shot	
	CoCoOp	RPO	CoCoOP	RPO	CoCoOp	RPO	CoCoOp	RPO	CoCoOp	RPO
Base	71.45±1.58	**71.69±0.30**	**73.93±1.26**	73.82±0.57	76.50±0.96	**77.18±0.71**	78.46±1.02	**79.66±0.36**	80.57±0.60	**81.31±0.30**
Novel	72.47±2.00	**73.82±0.73**	71.91±2.25	**73.83±0.64**	72.50±2.06	**73.43±0.67**	72.78±2.10	**73.66±0.50**	72.51±2.19	**75.47±0.25**
H.M	71.78±1.80	**72.69±0.37**	72.70±1.80	**73.77±0.45**	74.08±1.63	**75.05±0.45**	75.12±1.74	**76.27±0.28**	75.81±1.77	**78.11±0.10**

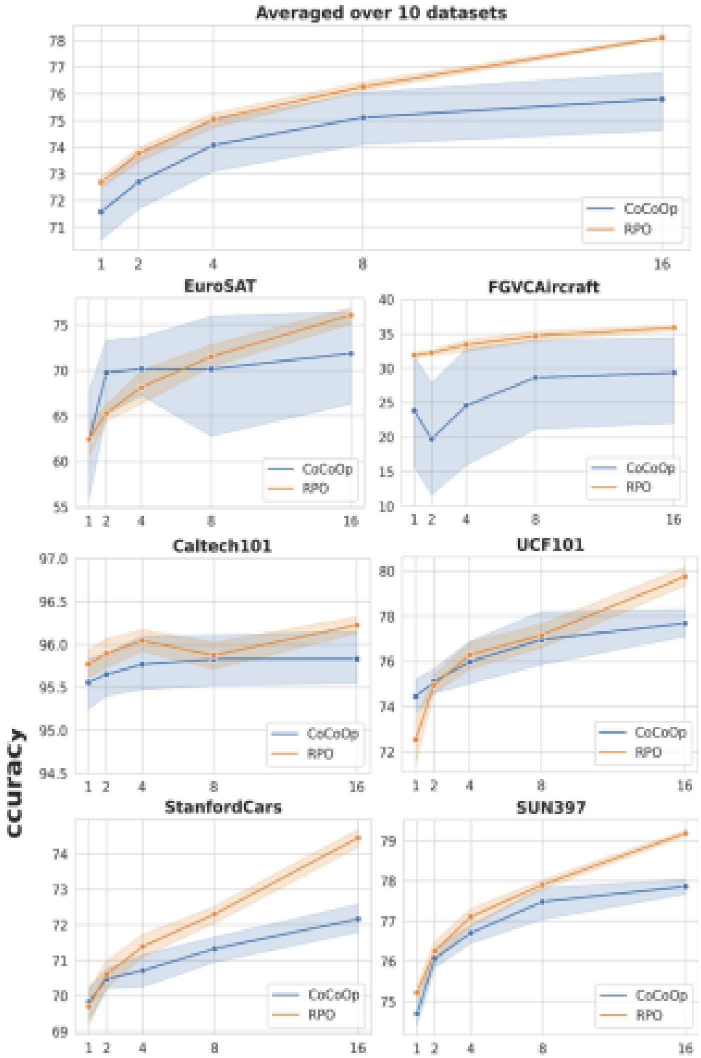
[Extreme few-shot setting]

Table 5: **Generalizability of uni-modal RPO.**

[Uni-modal prompts]

Methods	Base	Novel	H
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83
text-RPO	79.54	74.84	77.01
RPO	81.13	75.00	**77.78**

[Variance]



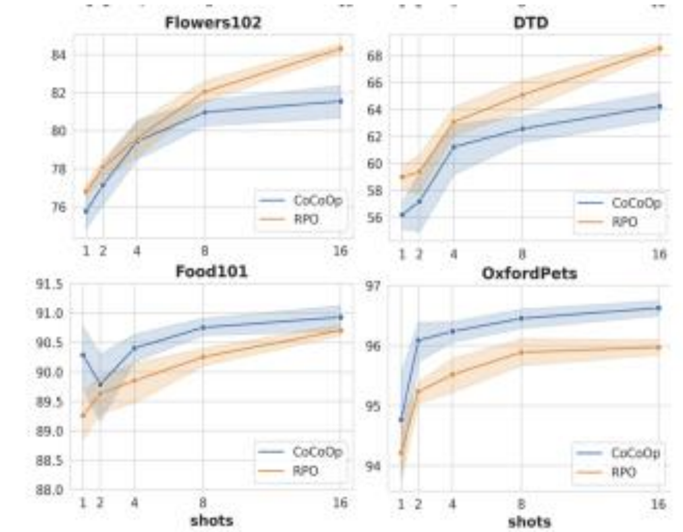


Figure 5: **Variance and generalization of RPO compared with CoCoOp.** RPO is more generalizable and robust than CoCoOp in the perspective of base to new generalization and lower performance variance.

Ⅴ. Conclusion

Proposed RPO, no internal representation shift, which results in better Generalization & Robustness
Initialized to special tokens
Good in base-to-new generalization and domain generalization with remarkably lower variance

Further research is needed to fully understand the efficiency and effectiveness of this method compared to other adaptation strategies

Ⅵ. Future works / Q & A

Thank you!