Adaptive Parameter Selection for Tuning Vision-Language Models

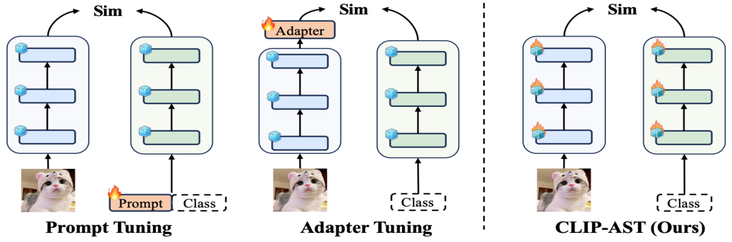
Yi Zhang[1*] Yi-Xuan Deng[3] Meng-Hao Guo[2] Shi-Min Hu[2†]
[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[2]BNRist, Department of Computer Science and Technology, Tsinghua University
[3] Zhili College, Tsinghua University

CVPR Nashville JUNE 11-15, 2025

Summary

Motivation: Improve model performance by fine-tuning the model's own parameters.

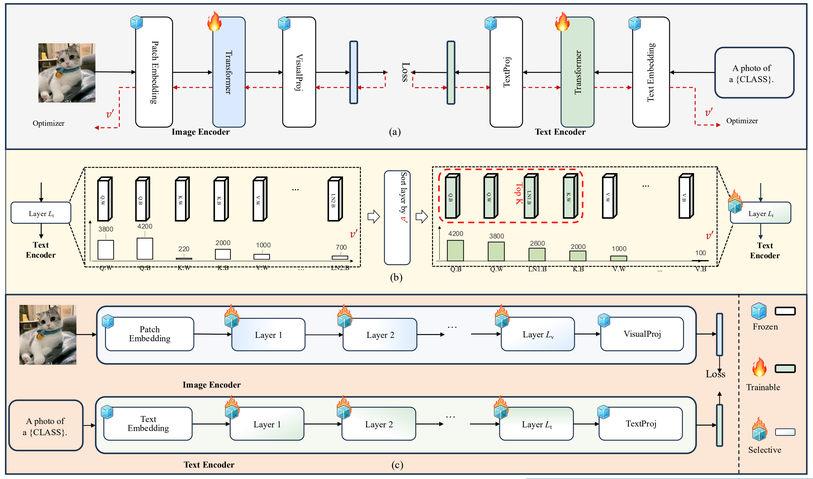
	Inference Efficiency	Training Efficiency	Adaptive Tuning Position	Performance	Adaptive Selective Tuning	Fixed Position Tuning/Freeze
Adapter	★★	★★★	★	★★		
Prompt	★★	★	★	★★		
Ours	★★★	★★	★★★	★★★		



Contribution:

- Propose a novel method that automatically selects and fine-tunes the most critical parameters in CLIP without extra parameters, maintaining the same inference efficiency as the original CLIP model.
- Introduce an automatic parameter identification strategy based on AdamW's adaptive learning rates to select important parameters.
- Outperforms CLIP and its variants on 13 datasets, achieving SOTA performance.

Method



(a) we first fine-tune the CLIP transformer block with sampling a batch of training data from the dataset.
(b) By counting the v_i' of each parameter in AdamW, we select the parameters that need to be fine-tuned.

$$\theta_i = (\theta_{i-1} - \alpha \cdot \lambda \cdot \theta_{i-1}) - \frac{\alpha}{\sqrt{\hat{v}_i} + \epsilon} \cdot \hat{m}_i \qquad v_i' = \mathrm{Avg}\left(\frac{1}{\sqrt{\hat{v}_i}}\right)$$

(c) Fine-tune the selected parameters using the following loss:

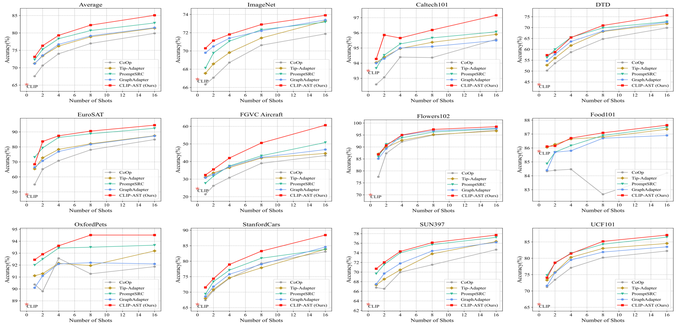
$$\mathcal{L}_{\text{SCL-image}} = \sum_{i=1}^{d} |\tilde{\mathbf{G}} - \mathbf{G}| \qquad \mathcal{L}_{\text{SCL-text}} = \sum_{i=1}^{d} |\tilde{\mathbf{W}} - \mathbf{W}|$$

$$\mathcal{L}_{\text{SCL-logits}} = \mathcal{D}_{\mathcal{KL}}(\mathrm{sim}(\tilde{\mathbf{G}}, \tilde{\mathbf{W}}), \mathrm{sim}(\mathbf{G}, \mathbf{W}))$$

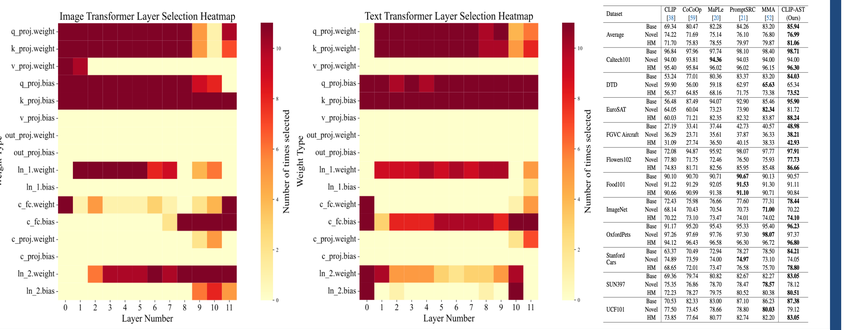
$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda_1 \cdot \mathcal{L}_{\text{SCL-image}} + \lambda_2 \cdot \mathcal{L}_{\text{SCL-text}} + \lambda_3 \cdot \mathcal{L}_{\text{SCL-logits}}$$

Experiments

Few-shots Classification Benchmark:



Heatmaps of parameter selection:



Base to new:

Dataset		CLIP [19]	CoCoOp [39]	MaPLe [20]	PromptSRC [21]	MMA [13]	CLIP-AST (Ours)
Average	Base	69.34	80.47	82.28	84.26	83.20	**85.94**
	Novel	74.22	71.69	75.14	76.10	76.80	**76.99**
	HM	71.70	75.83	78.55	79.97	79.87	**81.06**
Caltech101	Base	96.84	97.96	97.74	98.10	98.40	**98.71**
	Novel	94.00	93.81	**94.36**	94.03	94.10	94.00
	HM	95.40	95.84	96.02	96.02	96.15	**96.30**
DTD	Base	53.24	77.01	80.36	83.37	83.20	**84.83**
	Novel	59.90	56.00	59.18	62.97	**65.63**	65.34
	HM	56.37	64.85	68.16	71.75	73.38	**73.52**
EuroSAT	Base	56.48	87.49	94.07	92.90	85.46	**95.90**
	Novel	64.05	60.04	73.23	73.90	**82.34**	81.72
	HM	60.03	71.21	82.35	82.32	83.87	**88.24**
FGVC Aircraft	Base	27.19	33.41	37.44	42.73	40.57	**48.98**
	Novel	36.29	23.71	35.61	37.87	36.33	**38.21**
	HM	31.09	27.74	36.50	40.15	38.33	**42.93**
Flowers102	Base	72.08	94.87	95.92	98.07	97.77	**97.95**
	Novel	77.80	71.75	72.46	76.50	75.93	**77.53**
	HM	74.83	81.71	82.56	85.95	85.48	**86.66**
Food101	Base	90.10	90.70	90.71	**90.67**	90.13	90.57
	Novel	91.22	91.29	92.05	**91.53**	91.30	91.11
	HM	90.66	90.99	91.38	**91.10**	90.71	90.84
ImageNet	Base	72.43	75.98	76.66	77.60	77.31	**78.64**
	Novel	68.14	70.43	70.54	70.73	**71.00**	70.22
	HM	70.22	73.10	73.47	74.01	74.02	**74.10**
OxfordPets	Base	91.17	95.20	95.43	95.33	95.40	**96.23**
	Novel	97.26	97.69	97.76	97.30	**98.07**	97.37
	HM	94.12	96.43	96.58	96.30	96.72	**96.80**
StanfordCars	Base	63.37	70.49	72.94	78.27	78.50	**84.21**
	Novel	74.89	73.59	74.00	**74.97**	73.10	74.05
	HM	68.65	72.01	73.47	76.58	75.70	**78.80**
SUN397	Base	69.36	79.74	80.82	82.67	82.27	**83.05**
	Novel	75.35	76.86	78.70	78.47	**78.57**	78.12
	HM	72.23	78.27	79.75	80.52	80.38	**80.51**
UCF101	Base	70.53	82.33	83.00	87.10	86.23	**87.30**
	Novel	77.50	73.45	78.66	78.80	**80.03**	79.12
	HM	73.85	77.64	80.77	82.74	82.20	**83.05**