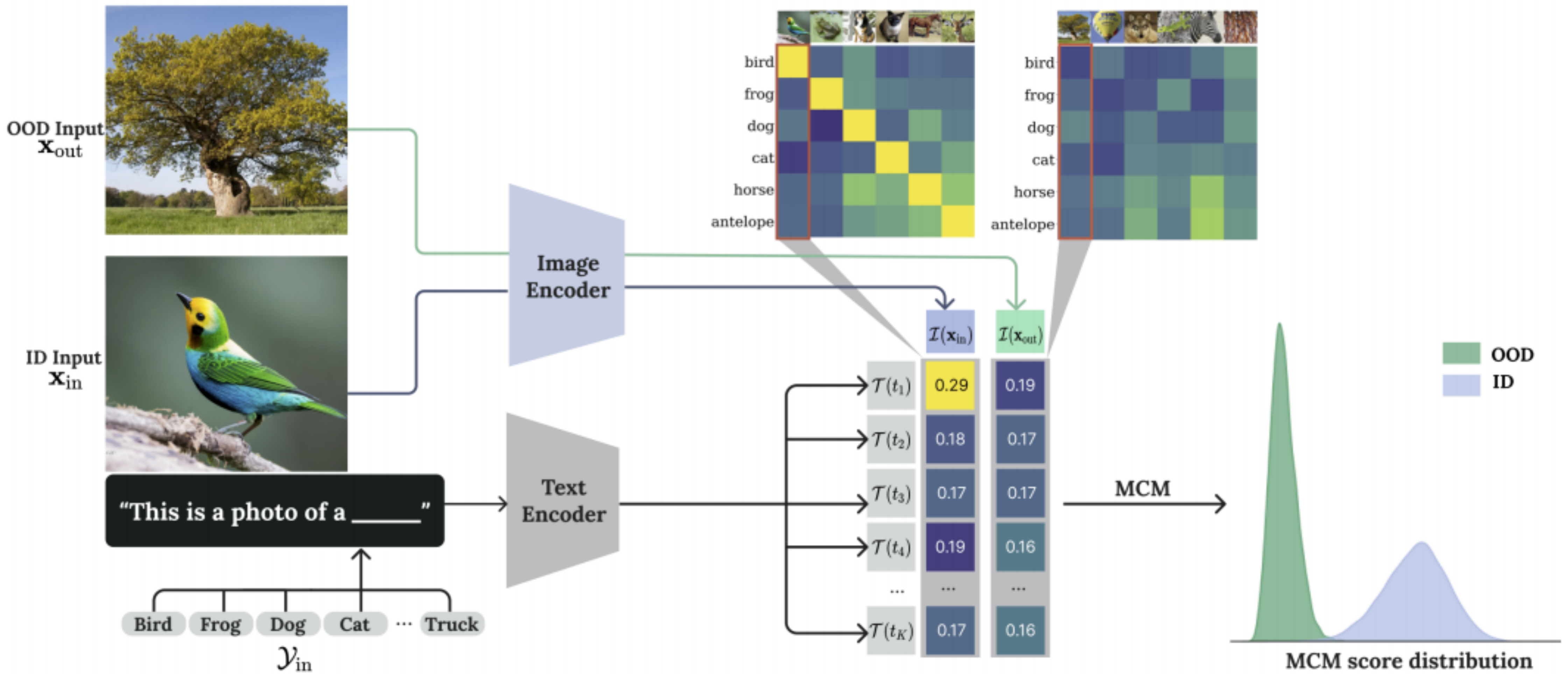
SeTAR: Out-of-Distribution Detection with Selective Low-Rank Approximation

Yixia Li[1*] Boya Xiong[2*] Guanhua Chen[1†] Yun Chen[2†]

[1]Southern University of Science and Technology
[2]Shanghai University of Finance Economics
[*]Equal Contribution [†]Corresponding Authors

Problem of OOD

- **ID (In-Distribution)**
  - Data specified by the user.
- **OOD (Out-of-Distribution)**
  - Data different from ID data set.
- **Target**
  - Maintain high accuracy on ID samples and effectively identify OOD samples.



$$G(\mathbf{x}) = \begin{cases} 1 & S(\mathbf{x}) \geq \lambda \\ 0 & S(\mathbf{x}) < \lambda \end{cases}$$

Related Work

- **CLIP-Based OOD Method [1][2]**
  - Utilizes global and local features within CLIP to measure image-concept alignment, enhancing the separation between ID and OOD samples.
  - Existing post-hoc methods yield from suboptimal performance.
- **CNN-Based OOD Post-Hoc Method [3][4]**
  - Assumes that ID and OOD samples produce distinct activation patterns in models trained on ID data. Rectifying activations can reduce OOD influence and improve ID-OOD separability.
  - The CLIP model is not fine-tuned on downstream ID-domain datasets.
  - Activation differences between ID and OOD data become more subtle.

[1] https://openreview.net/forum?id=KnCS9390Va
[2] https://arxiv.org/abs/2304.04521
[3] https://openreview.net/forum?id=ndYXTEL6cZz
[4] https://arxiv.org/abs/2111.12797

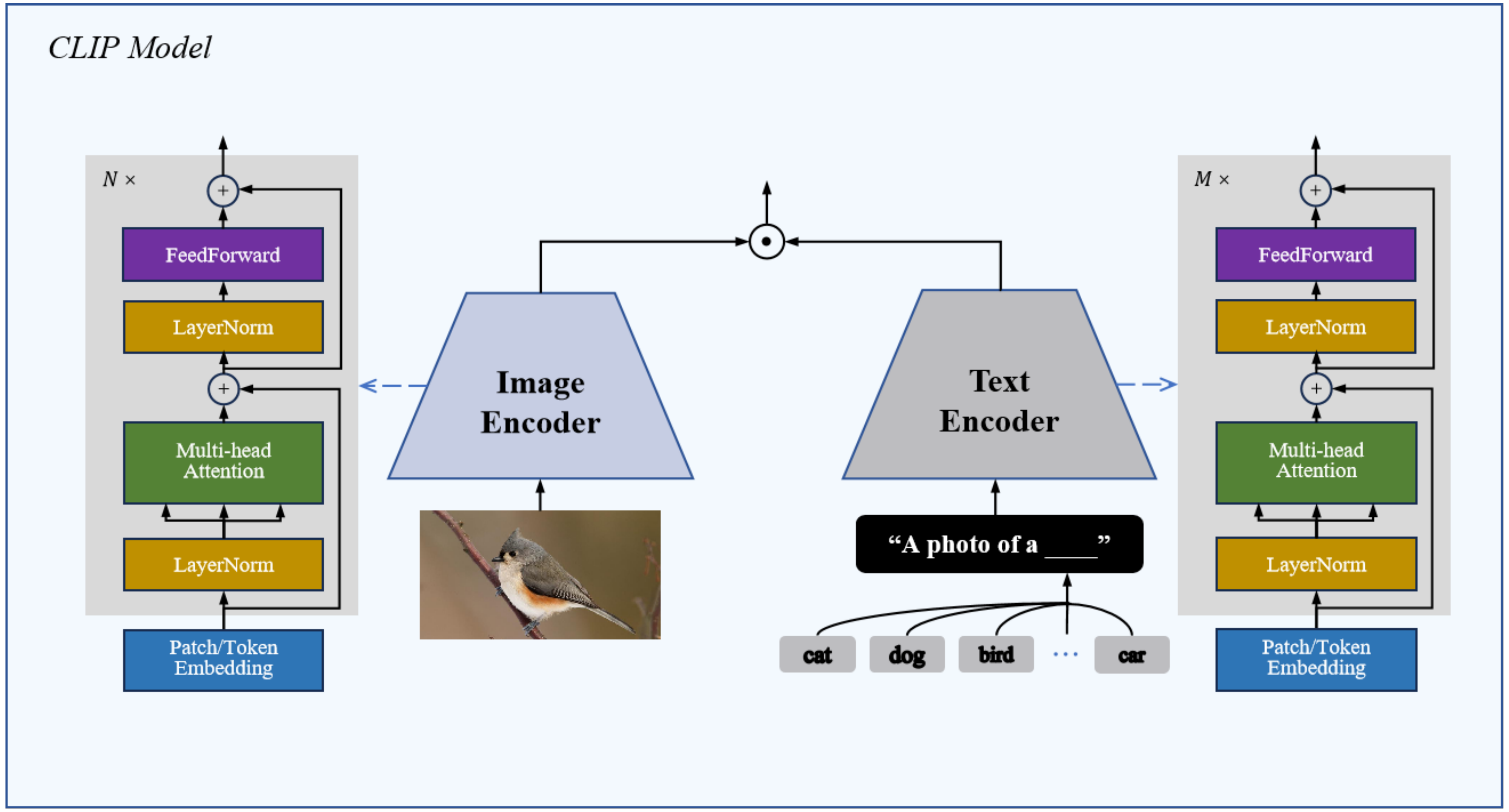
Motivation

- **Low-Rank Approximations Improve Stability [5]**
  - Minor singular components often **contain noisy information** that can compromise stability and lead to incorrect model responses.
  - Using low-rank approximations in certain layers of transformer blocks can enhance LLMs' question-answering performance.
- **Our Method**
  - Selectively reduce the rank of model weights, prioritizing crucial information while discarding noisy elements.

[5] https://arxiv.org/abs/2312.13558

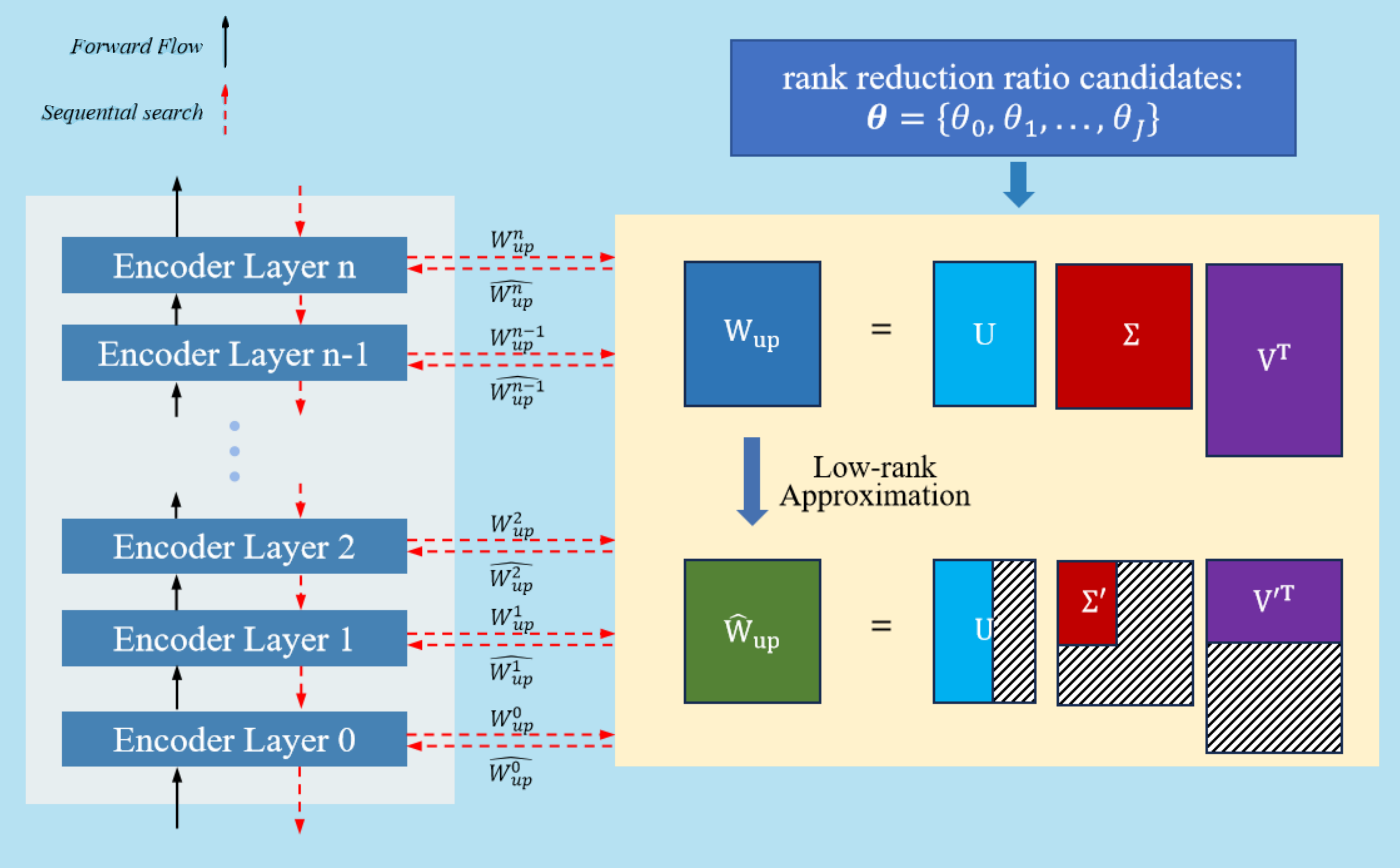
Method - Training Free

- **SeTAR - Training Free**
  - Given a CLIP model.
  - Define a list of rank reduction ratio candidates: $\boldsymbol{\theta} = \{\theta_0, \ \theta_1, \ \ldots, \ \theta_J\}$.
  - Optimize the rank reduction with top-to-bottom, image-to-text greedy search algorithm.



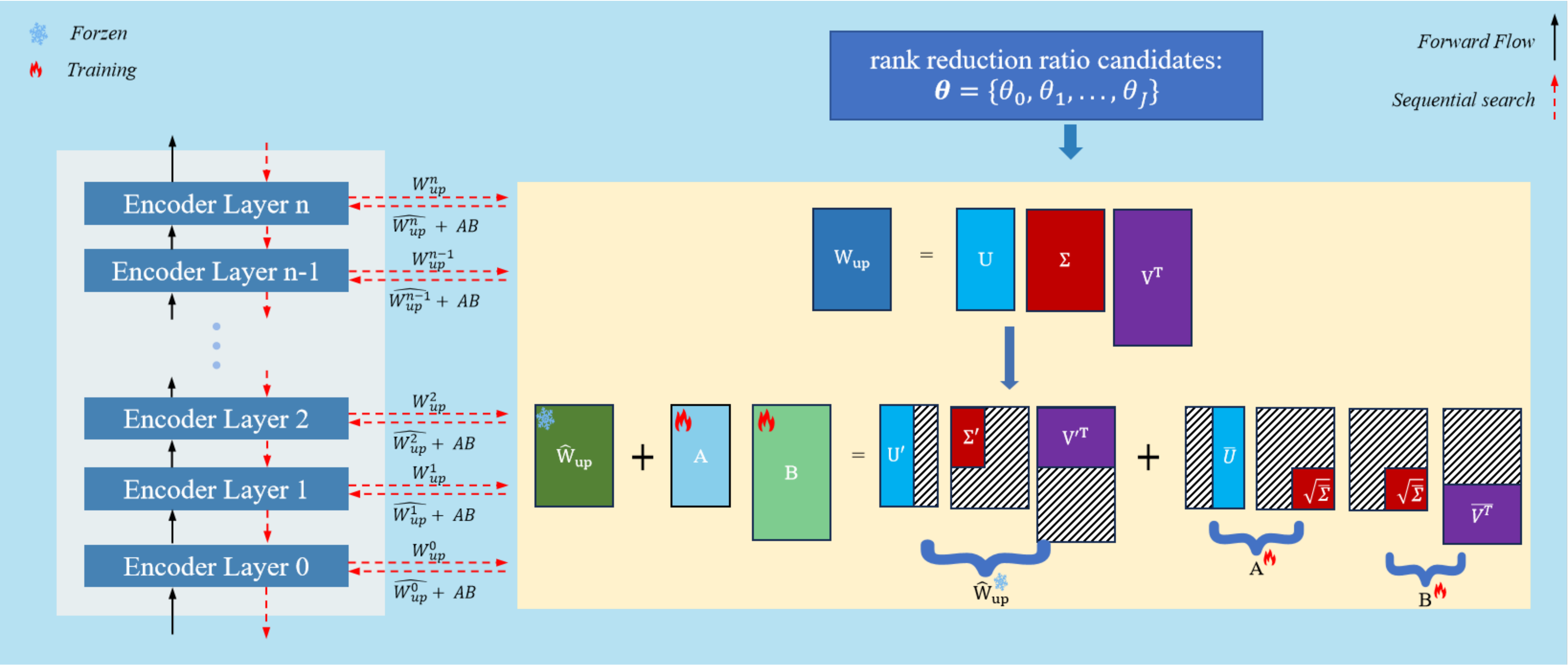
Method - Training Free

- **How to Process**
  - Searching for the optimal W_{up}^{n} ratio using metrics.
  - Replacing W_{up}^{n} with its low-rank approximation ($\hat{W}_{up}^{n}$).
  - Sequentially process all N image encoder layers from top to bottom.
  - Same process for text encoder.



Method - Fine Tuning

- **SeTAR - Fine Tuning**



$$W = \widehat{W} + B \times A$$

$$B = \sum_{i=r+1}^{\min(m,n)} \sqrt{\sigma_i^{\downarrow}(W)} u_i$$

$$A = \sum_{i=r+1}^{\min(m,n)} \sqrt{\sigma_i^{\downarrow}(W)} v_i^{\top}$$

Expriments

• SeTAR - Training Free

Table 1: Training free results compared to zero-shot baselines on CLIP-base. **Bold** value represent the highest performance.

Method	iNaturalist		SUN		Places		Texture		ImageNet22K		COCO		**Average**	
	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
ImageNet1K														
MCM Score														
Vanilla MCM†	30.91	94.61	37.59	92.57	44.69	89.77	57.77	86.11	-	-	-	-	42.74	90.77
Vanilla MCM*	32.07	94.43	38.65	92.37	43.73	90.03	57.89	86.13	-	-	-	-	43.09	90.74
SeTAR	**26.92**	**94.67**	**35.57**	**92.79**	**42.64**	**90.16**	**55.83**	**86.58**	-	-	-	-	**40.24**	**91.05**
GL-MCM Score														
Vanilla GL-MCM†	15.18	96.71	30.42	93.09	38.85	89.90	57.93	83.63	-	-	-	-	35.47	90.83
Vanilla GL-MCM*	15.34	96.62	30.65	93.01	37.76	90.07	57.41	83.73	-	-	-	-	35.29	90.86
SeTAR	**13.36**	**96.92**	**28.17**	**93.36**	**36.80**	**90.40**	**54.17**	**84.59**	-	-	-	-	**33.12**	**91.32**
Pascal-VOC														
MCM Score														
Vanilla MCM†	8.20	98.23	28.60	94.68	◇	◇	51.70	91.45	51.40	90.94	54.50	89.02	38.88	92.86
Vanilla MCM*	7.24	98.23	27.91	94.56	32.40	92.45	51.61	91.89	50.60	91.42	53.70	89.30	37.24	92.98
SeTAR	**4.59**	**98.71**	**24.91**	**95.15**	**28.46**	**93.21**	**40.44**	**93.58**	**48.25**	**92.08**	**48.10**	**89.70**	**32.46**	**93.74**
GL-MCM Score														
Vanilla GL-MCM†	4.20	98.71	23.10	94.66	◇	◇	43.00	92.84	41.00	92.38	44.30	90.48	31.12	93.81
Vanilla GL-MCM*	4.33	98.81	22.94	94.63	26.20	93.11	41.61	92.88	37.88	93.17	43.70	90.71	29.44	93.88
SeTAR	**3.66**	**98.96**	**21.93**	**94.81**	**25.04**	**93.62**	**20.35**	**96.36**	**31.47**	**94.31**	**40.70**	**91.19**	**23.86**	**94.87**

Expriments

- **SeTAR - Fine Tuning**

Table 2: Fine-tuning results on ImageNet1K benchmark. **Bold** values indicate the highest performance. ± indicates the standard deviation from 3 runs.

CLIP-base	MCM Score FPR95↓	MCM Score AUROC↑	GL-MCM Score FPR95↓	GL-MCM Score AUROC↑
NPOS†	42.20	90.43	36.86	90.37
CoOp†	44.81	90.03	36.58	90.25
LoCoOp†	40.17	91.53	33.52	92.14
LoCoOp*	$39.76_{\pm 4.06}$	$91.22_{\pm 0.52}$	$34.14_{\pm 1.64}$	$91.73_{\pm 0.17}$
LoRA*	$41.67_{\pm 0.14}$	$90.85_{\pm 0.01}$	$34.36_{\pm 0.11}$	$90.88_{\pm 0.01}$
SeTAR+FT	$\mathbf{38.77}_{\pm 0.22}$	$\mathbf{91.55}_{\pm 0.01}$	$\mathbf{32.19}_{\pm 0.20}$	$\mathbf{92.31}_{\pm 0.05}$
CLIP-large	MCM Score FPR95↓	MCM Score AUROC↑	GL-MCM Score FPR95↓	GL-MCM Score AUROC↑
LoCoOp*	$40.74_{\pm 3.80}$	$91.13_{\pm 0.79}$	$46.74_{\pm 4.19}$	$89.32_{\pm 0.80}$
LoRA*	$38.62_{\pm 0.07}$	$91.66_{\pm 0.02}$	$43.39_{\pm 0.01}$	$89.76_{\pm 0.03}$
SeTAR+FT	$\mathbf{34.75}_{\pm 0.55}$	$\mathbf{92.86}_{\pm 0.15}$	$\mathbf{37.05}_{\pm 0.59}$	$\mathbf{91.83}_{\pm 0.12}$
Swin-base	MSP Score FPR95↓	MSP Score AUROC↑	Energy Score FPR95↓	Energy Score AUROC↑
LoRA*	$57.02_{\pm 0.03}$	$80.49_{\pm 0.01}$	$62.17_{\pm 0.02}$	$72.80_{\pm 0.00}$
SeTAR+FT	$\mathbf{47.12}_{\pm 0.42}$	$\mathbf{87.80}_{\pm 0.44}$	$\mathbf{39.29}_{\pm 0.57}$	$\mathbf{88.01}_{\pm 0.51}$

Ablation Experiment

- **Different Backbones**
  - Compatible with diverse model backbones.
- **Various Score Functions**
  - Consistently outperforms baselines across all backbones with various scoring functions.
- **Classification Performance**
  - Maintains or improves classification accuracy.

Table 3: Results for different ViT backbones

Backbone	Score	Vanilla Method		SeTAR	
		FPR↓	AUC↑	FPR↓	AUC↑
ImageNet1K					
CLIP-base	NegLabel	25.40	94.21	**23.09**	**94.48**
CLIP-large	MCM	37.19	91.73	**36.26**	**91.92**
CLIP-large	GL-MCM	40.65	89.98	**39.54**	**90.22**
Swin-base	MSP	59.25	84.12	**56.05**	**85.77**
Swin-base	Energy	65.01	76.10	**51.61**	**84.42**
Pascal-VOC					
CLIP-large	MCM	52.21	91.68	**42.57**	**92.91**
CLIP-large	GL-MCM	43.96	92.45	**31.12**	**94.00**

Table 4: Image classification results

Method	IN1K	SUN	Places	Texture	**Average**
Vanilla CLIP*	64.07	75.77	45.65	43.60	57.27
LoCoOp*	64.93	75.89	46.47	37.79	56.27
LoRA*	65.43	76.86	46.58	**43.98**	58.21
SeTAR	63.97	75.50	45.81	43.76	57.26
SeTAR+FT	**67.02**	**77.94**	**46.64**	43.28	**58.72**

Conclusion

- SeTAR is a simple, effective and training-free OOD detection method using post-hoc low-rank approximation.
- SeTAR scales across unimodal and multimodal models, enhancing various scoring functions.
- We extend SeTAR to SeTAR+FT, a finetuning approach that adapts models to in-distribution data, achieving state-of-the-art OOD detection.

Thanks for listening

Paper

Code

Project

Homepage