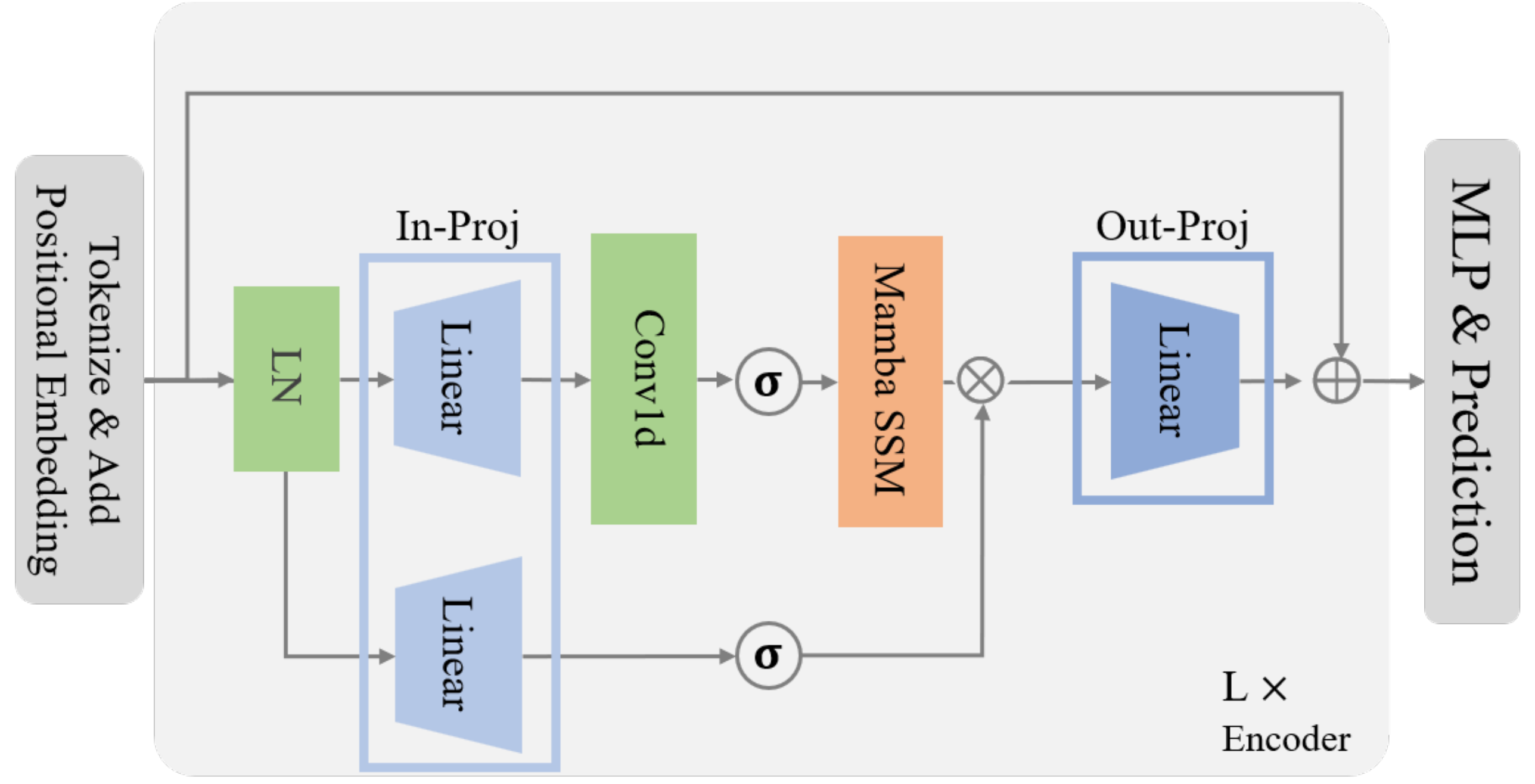
Parameter Efficient Mamba Tuning via Projector-targeted Diagonal-centric Linear Transformation

Seokil Ham, Hee-Seon Kim, Sangmin Woo, Changick Kim
Korea Advanced Institute of Science and Technology (KAIST)

Mamba Architecture

- Mamba architecture introduces **selective SSMs** and **hardware-aware operations** for dynamic and linear-time computation with respect to input size.
- Both Mamba LLM and Vision Mamba share the block below.

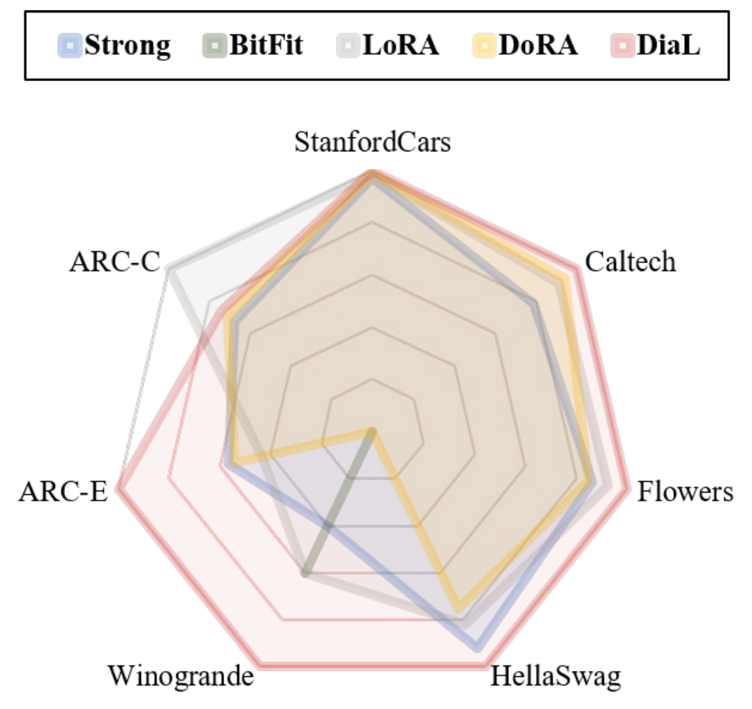


Motivation

- PEFT methods for Mamba remain largely unexplored.
- We observe that **Projectors play a key role** in Mamba PEFT.

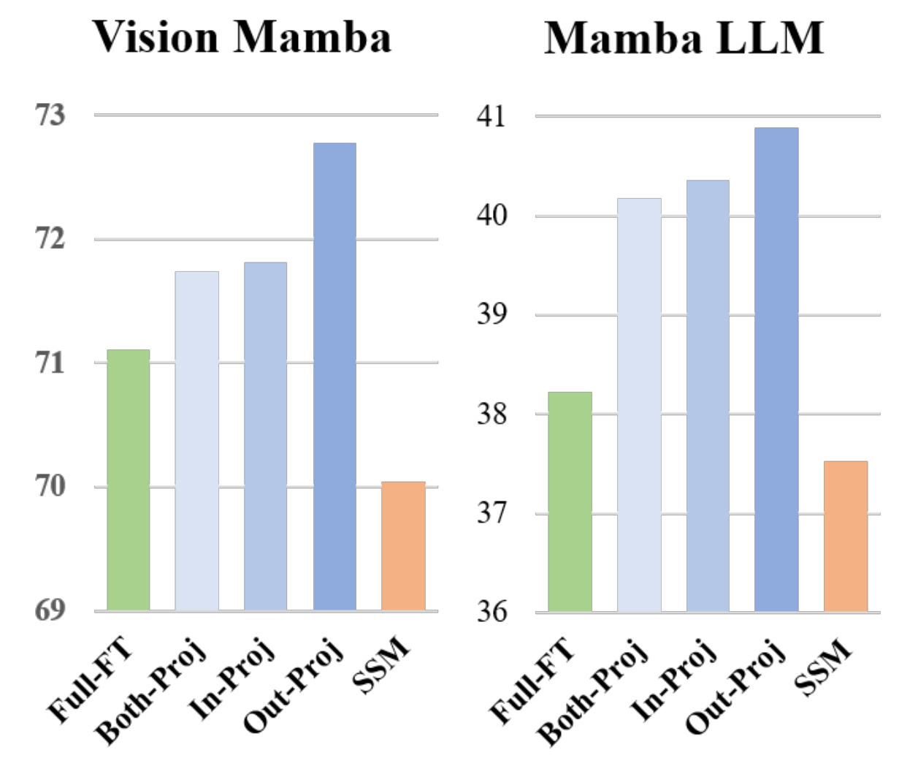
Contributions

- First investigation of Projectors in Mamba architecture.
- Based on our analysis of projectors, we propose **ProDiaL**, the first projector-targeted PEFT method for Mamba.
- Experiments on both Mamba LLM and Vision Mamba show that applying ProDiaL and other PEFT methods to projectors significantly outperforms targeting other components.



Observations

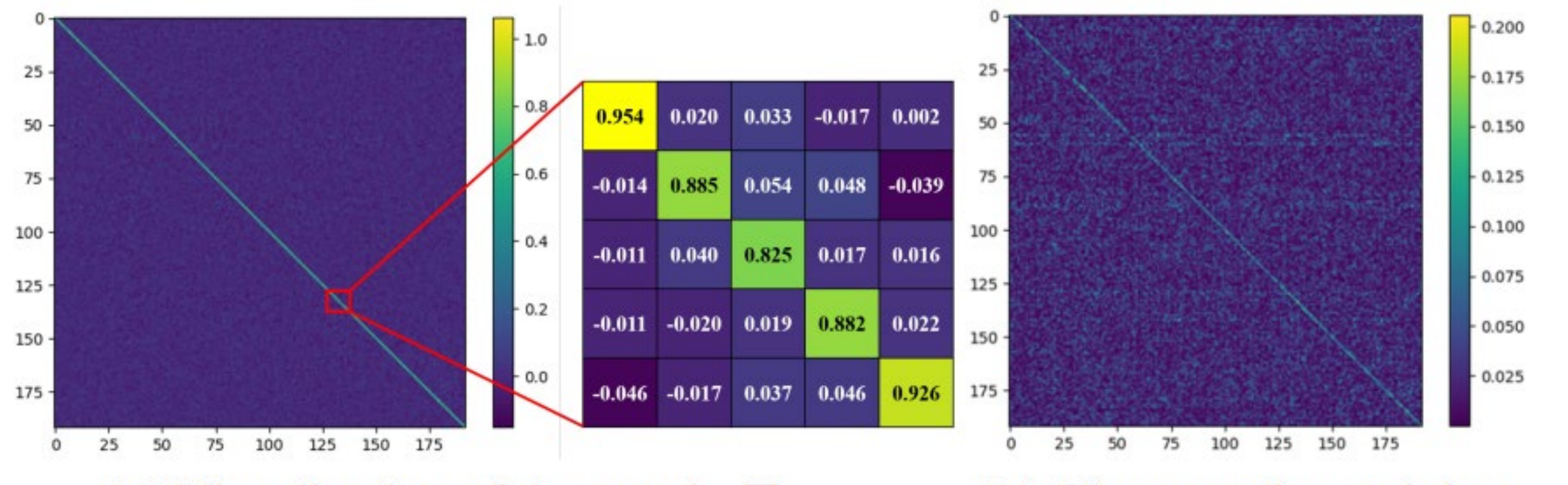
- **Projectors play a more critical role than SSMs** in learning downstream task knowledge.



- **Diagonal Entries in T are dominant.**

$$W' = WT,$$
$$T_{det} = W^{-1}W',$$

W': Fine-tuned Projector, W: Pre-trained Projector



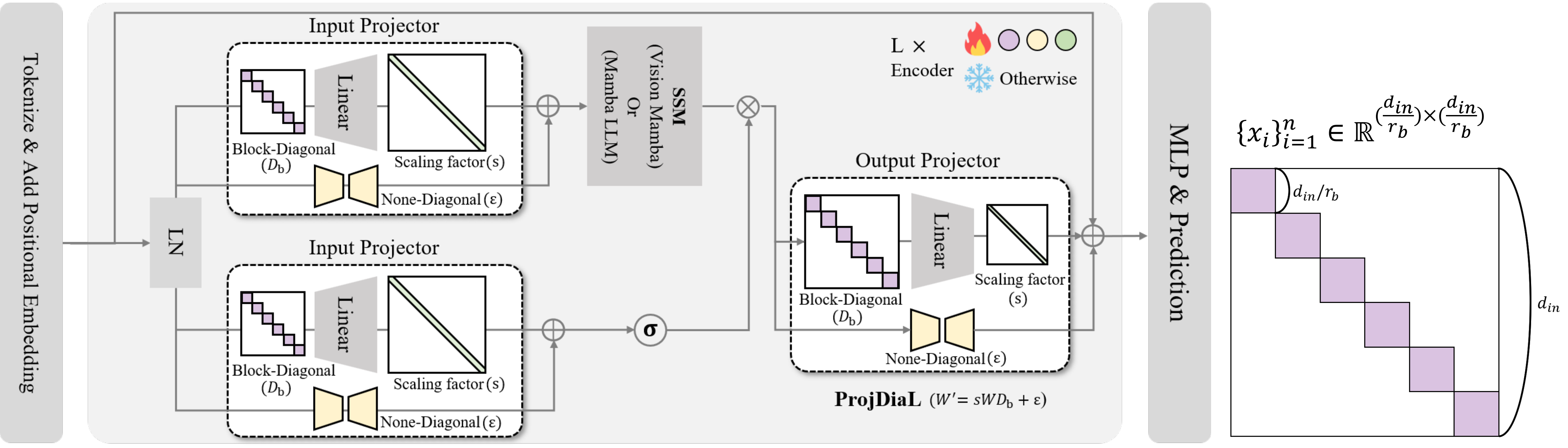
(a) Visualization of the matrix T_{det}. (b) Changes after training.

Method: ProDiaL

- We propose a novel Mamba Projector-targeted PEFT method, **ProDiaL** (**Pro**jector-targeted **Dia**gonal-centric **L**inear Transformation).
- ProDiaL decomposes T into Diagonal and Off-diagonal Entries, and trains them separately.

$$W' = WT = sWD_b + \epsilon,$$
$$D_b = [\mathbb{I} - relu(\mathbb{I} * D_a)] + (\mathbf{1} - \mathbb{I}) * D_a,$$
$$D_a = diag(x_1, x_2, \ldots, x_n),$$
$$\epsilon = B_\epsilon A_\epsilon$$

s: learnable scaling parameter, A_ϵ and B_ϵ are low-rank matrices.



Experiment Results

- Performance on Mamba1 architecture.

		Mamba LLM					Vision Mamba			
	Method	HellaSwag	Winogrande	ARC-E	ARC-C	Avg	StanfordCars	Caltech	Flowers	Avg.
Baselines	Full-FT	38.23 (130.00M)	53.12 (130.00M)	53.54 (130.00M)	28.84 (130.00M)	43.43	90.06 (7.00M)	92.86 (7.00M)	92.05 (7.00M)	91.66
	Linear Probing	-	-	-	-	-	57.46 (0.04M)	91.10 (0.02M)	59.90 (0.02M)	69.49
	BitFit [49]	35.69 (0.07M)	53.12 (0.07M)	52.86 (0.07M)	26.88 (0.07M)	42.14	65.51 (0.08M)	93.71 (0.06M)	78.84 (0.06M)	79.35
	Strong [17]	38.66 (3.80M)	53.04 (3.80M)	54.17 (3.80M)	28.67 (3.80M)	43.64	84.78 (0.96M)	95.70 (0.94M)	86.76 (0.94M)	89.08
Both-Proj	FT	40.18 (84.94M)	52.57 (84.94M)	54.38 (84.94M)	29.52 (84.94M)	44.16	89.67 (5.35M)	95.01 (5.33M)	92.00 (5.33M)	92.22
	LoRA	38.33 (2.36M)	53.12 (2.36M)	53.87 (2.36M)	**29.52** (2.36M)	43.71	85.06 (0.63M)	96.01 (0.61M)	87.32 (0.61M)	89.46
	DoRA	38.13 (2.45M)	52.88 (2.45M)	54.12 (2.45M)	28.75 (2.45M)	43.47	85.18 (0.69M)	96.09 (0.65M)	86.60 (0.65M)	89.29
	ProDiaL	**38.92** (2.42M)	**53.28** (2.42M)	**55.18** (2.38M)	28.84 (2.38M)	**44.06**	**85.38** (0.67M)	**96.24** (0.65M)	**88.00** (0.65M)	**89.87**
In-Proj	FT	40.36 (56.62M)	53.20 (56.62M)	54.59 (56.62M)	29.61 (56.62M)	44.44	89.62 (3.58M)	95.24 (3.56M)	91.02 (3.56M)	91.96
	LoRA	**38.46** (1.48M)	52.80 (1.48M)	53.87 (1.48M)	28.41 (1.48M)	43.39	82.12 (0.41M)	95.78 (0.39M)	85.71 (0.39M)	87.87
	DoRA	38.08 (1.55M)	52.64 (1.55M)	54.04 (1.55M)	28.50 (1.55M)	43.32	82.17 (0.43M)	95.55 (0.41M)	85.95 (0.41M)	87.89
	ProDiaL	38.41 (1.49M)	**52.96** (1.49M)	**54.50** (1.25M)	**29.61** (1.25M)	**43.87**	**82.45** (0.42M)	**95.93** (0.41M)	**85.97** (0.33M)	**88.12**
Out-Proj	FT	40.89 (28.31M)	52.80 (28.31M)	55.09 (28.31M)	29.27 (28.31M)	44.51	88.86 (1.81M)	95.63 (1.77M)	91.45 (1.77M)	91.98
	LoRA	37.30 (0.89M)	53.12 (0.89M)	53.66 (0.89M)	28.54 (0.89M)	43.08	77.81 (0.26M)	95.40 (0.24M)	80.60 (0.24M)	84.60
	DoRA	37.19 (0.90M)	52.88 (0.90M)	53.66 (0.90M)	28.67 (0.90M)	43.10	77.70 (0.28M)	95.47 (0.26M)	80.97 (0.26M)	84.71
	ProDiaL	**38.19** (0.92M)	**53.75** (0.92M)	**54.84** (0.90M)	**30.80** (0.90M)	**44.40**	**78.00** (0.27M)	**95.55** (0.25M)	**81.90** (0.25M)	**85.15**

- Performance on Mamba2 architecture.

	Method	HellaSwag	Winogrande	ARC-E	ARC-C	Avg
	Full-FT	38.23 (130.00M)	53.12 (130.00M)	53.54 (130.00M)	28.84 (130.00M)	43.43
Both-Proj	FT	38.76 (90.1M)	53.12 (90.1M)	50.67 (90.1M)	28.84 (90.1M)	42.84
	LoRA	38.50 (2.47M)	53.35 (2.47M)	52.36 (2.47M)	30.20 (2.47M)	43.60
	DoRA	35.24 (2.57M)	52.01 (2.57M)	47.18 (2.57M)	24.15 (2.57M)	39.65
	ProDiaL	**38.57** (2.44M)	**53.83** (2.00M)	**53.03** (2.33M)	**30.46** (2.22M)	**43.97**
In-Proj	FT	39.89 (61.8M)	53.35 (61.8M)	51.56 (61.8M)	27.22 (61.8M)	43.01
	LoRA	37.32 (1.58M)	53.43 (1.58M)	52.02 (1.58M)	30.03 (1.58M)	43.20
	DoRA	35.24 (1.66M)	52.01 (1.66M)	47.18 (1.66M)	24.15 (1.66M)	39.65
	ProDiaL	**37.91** (0.98M)	**53.75** (0.89M)	**53.03** (0.98M)	**30.29** (0.93M)	**43.75**
Out-Proj	FT	40.62 (28.3M)	53.43 (28.3M)	54.08(28.3M)	29.10 (28.3M)	44.31
	LoRA	37.44 (0.89M)	53.28 (0.89M)	52.65 (0.89M)	28.50 (0.89M)	42.97
	DoRA	37.44 (0.90M)	53.59 (0.90M)	52.69 (0.90M)	28.92 (0.90M)	43.16
	ProDiaL	**37.86** (0.90M)	**53.51** (0.50M)	**53.45** (0.68M)	**30.29** (0.90M)	**43.78**

- Scalability of our ProDiaL

	Method	Mamba-370M	Mamba-1.4B	Vim-small
Base	Full-FT	56.99 (370M)	61.17 (1.40B)	94.09 (25.45M)
	BitFit [49]	56.99 (0.20M)	61.25 (0.39M)	96.62 (0.11M)
	Strong [17]	57.14 (8.19M)	61.80 (0.06B)	96.47 (1.85M)
Both-Proj	FT	57.22 (302M)	61.72 (1.21B)	94.94 (21.27M)
	LoRA	56.75 (6.29M)	61.72 (0.05B)	96.85 (1.22M)
	DoRA	56.99 (6.54M)	61.72 (0.05B)	96.85 (1.27M)
	ProDiaL	**57.06** (5.75M)	**61.96** (0.05B)	**97.16** (1.32M)
In-Proj	FT	57.06 (201M)	61.56 (0.81B)	95.17 (14.20M)
	LoRA	56.75 (3.93M)	**61.56** (0.03B)	**97.16** (0.88M)
	DoRA	**57.22** (4.13M)	61.48 (0.03B)	**97.16** (0.91M)
	ProDiaL	**57.22** (3.74M)	**61.56** (0.03B)	97.09 (0.82M)
Out-Proj	FT	57.22 (101M)	61.72 (0.40B)	95.86 (7.12M)
	LoRA	56.91 (2.36M)	61.56 (0.02B)	96.70 (0.48M)
	DoRA	57.22 (2.41M)	61.48 (0.02B)	96.78 (0.59M)
	ProDiaL	**57.30** (2.02M)	**61.80** (0.02B)	**96.85** (0.51M)