



Background & Motivation

- Grounded on large-scale pre-trained models, meta-tuning helps models quickly adapt to new tasks in few-shot scenarios.
- Meta-tuning on single domain yields **marginal OOD improvements** over pre-trained models.
- Meta-tuning suffers from **vulnerability in adversarial attacks and common visual corruptions** under distribution shifts.

Contributions

- We propose **AMT**, a novel **adversarial meta-tuning** approach for enhancing the robust generalization of pre-trained vision transformers across diverse domains.
- We construct the **adaptive robust LoRAPool** by injecting the adversarial perturbations on the **inputs, singular values and vectors of the weight matrices** under varying perturbation budgets during meta-tuning.
- The discriminative components of the pool are integrated into the pre-trained model via a simple yet effective **test-time merging mechanism** for task adaptation.

Adaptive Robust LoRAPool

- Generate Adversarial Query Set:** Use PGD to generate adversarial query images with different robustness strength.

$$\begin{aligned} & \max_{\|\delta\|_{\infty} \leq \epsilon_i} \mathcal{L}(f_{\theta}(S, x_q + \delta), y_q) \\ & \delta \leftarrow \Pi_{\epsilon_i} \left(\delta + \alpha \cdot \text{sign} \left(\nabla_{\delta} \mathcal{L}(f_{\theta}(S, x_q + \delta), y_q) \right) \right) \end{aligned} \quad (1)$$

- Adversarial Perturbation on Singular Values and Vectors**

Initialize LoRA parameters with the SVD results and freeze the residual part.

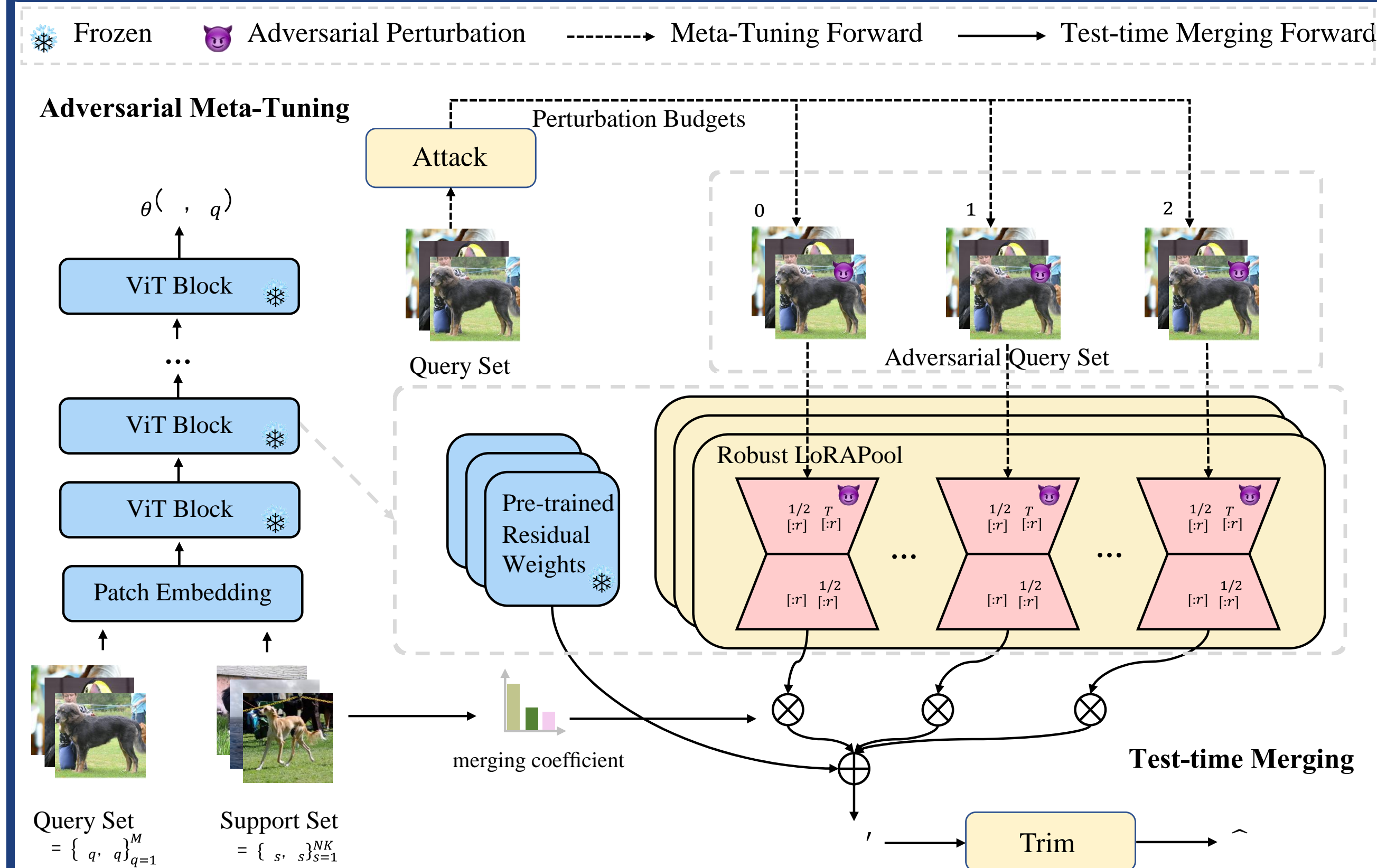
$$\begin{aligned} A &= U_{[r]} \text{diag}(S_{[r]}^{1/2}) \\ B &= \text{diag}(S_{[r]}^{1/2}) V_{[r]}^T \end{aligned} \quad (2)$$

$$W^{\text{res}} = U_{[r]} \text{diag}(S_{[r]}) V_{[r]}^T$$

Incorporate worst-case perturbation on A and B using gradient ascent.

$$\begin{aligned} \delta_A &= \eta_1 \cdot \frac{1}{M} \sum_{q=1}^M \nabla_A \mathcal{L}(f_{W^{\text{res}} + AB}(S, x_q^{\text{adv}}), y_q) \\ A &\leftarrow A - \eta_2 \cdot \frac{1}{M} \sum_{q=1}^M \nabla_A \mathcal{L}(f_{W^{\text{res}} + (A + \delta_A)B}(S, x_q^{\text{adv}}), y_q) \end{aligned} \quad (3)$$

Framework



Pseudo-code

Algorithm 1 Robust LoRAPools

- Input:** Source training domain D_{tr}^{seen} , pre-trained weight residual matrix W^{res} ; P sets of attack configuration candidates;
- Output:** Adversarially meta-trained LoRAPool;
- Initialize adversarial LoRAPool: $\phi = \{\}$
- for** $i = 1$ to P (in parallel) **do**
- Sample the i -th set of ϵ_i, α_i from attack configuration candidates.
- Initialize the LoRA parameter AB via Eq. (2);
- while** not converged **do**
- Generate adversarial query set $\mathcal{Q}_{adv} = \{x_q^{\text{adv}}, y_q\}_{q=1}^M$ with ϵ_i, α_i via Eq. (1)
- // Perturb singular value and vectors**
- $\delta_A = \eta_1 \cdot \frac{1}{M} \sum_{q=1}^M \nabla_A \mathcal{L}(f_{W^{\text{res}} + AB}(S, x_q^{\text{adv}}), y_q)$
- $\delta_B = \eta_1 \cdot \frac{1}{M} \sum_{q=1}^M \nabla_B \mathcal{L}(f_{W^{\text{res}} + AB}(S, x_q^{\text{adv}}), y_q)$
- // Update AB via SGD**
- $A \leftarrow A - \eta_2 \cdot \frac{1}{M} \sum_{q=1}^M \nabla_A \mathcal{L}(f_{W^{\text{res}} + (A + \delta_A)B}(S, x_q^{\text{adv}}), y_q)$
- $B \leftarrow B - \eta_2 \cdot \frac{1}{M} \sum_{q=1}^M \nabla_B \mathcal{L}(f_{W^{\text{res}} + (A + \delta_A)B}(S, x_q^{\text{adv}}), y_q)$
- end while**
- $\phi = \phi \cup AB$
- end for**

Algorithm 2 Test-Time Merging

- Input:** Support set of meta-test task $\mathcal{S} = \{x_s^s, y_s^s\}_{s=1}^{NK}$, pre-trained residual weight matrix W^{res} , adaptive robust LoRAPool $\phi = [A_1 B_1, \dots, A_P B_P]$
- for** $i = 1, \dots, P$ (in parallel) **do**
- // Calculate the intra-class compactness**
- $C_i = \frac{1}{NK} \sum_{s=1}^{NK} \gamma(f_{W^{\text{res}} + A_i B_i}(x_s), \mathbf{p}_{y_s})$
- // Calculate the inter-class divergence**
- $V_i = \frac{1}{NK} \sum_{s=1}^{NK} \sum_{c \neq y_s} \gamma(f_{W^{\text{res}} + A_i B_i}(x_s), \mathbf{p}_c)$
- end for**
- $\zeta_i = \frac{\text{Top}_k(\exp(-\beta(1 - (\lambda C - (1 - \lambda)V)))_i)}{\sum_{i=1}^P \text{Top}_k(\exp(-\beta(1 - (\lambda C - (1 - \lambda)V)))_i)}$
- $W' = W^{\text{res}} + \sum_{i=1}^P \zeta_i A_i B_i$
- // Singular Value Trimming**
- $\hat{W} = \text{trim}(W')$

Experiments

Table 1. Few-shot classification clean accuracy (%) on Meta-Dataset in the 5-way 1-shot setting.

1-shot	Backbone	TTF	In-domain		Out-of-domain										Avg.
			ImageNet	Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO			
PM [12]	ViT-small	-	65.07	59.03	38.13	76.18	61.56	57.29	56.03	80.41	55.17	54.42	60.35		
StyleAdv [86]	ViT-small	-	56.10	62.25	40.38	66.62	55.94	57.93	53.19	81.10	54.20	48.08	57.58		
AMT	ViT-small	-	68.80	71.95	42.90	79.95	62.99	59.62	59.06	85.37	63.78	57.14	65.16		
PMF [12]	ViT-small	Y	65.07	71.52	38.67	76.15	61.62	59.82	56.03	80.41	59.71	54.41	62.34		
PMF+AMT-FT	ViT-small	Y	68.80	77.83	42.90	79.95	63.77	63.72	59.06	85.37	63.87	57.37	66.26		
ATTNSCALE [59]	ViT-small	Y	63.66	72.51	40.09	73.59	61.04	60.26	54.88	82.52	59.91	55.10	62.36		
ATTNSCALE+AMT-FT	ViT-small	Y	68.80	79.43	42.90	79.95	63.08	65.66	59.06	85.37	64.13	58.24	66.66		

Table 2. Few-shot classification clean accuracy (%) on BSCD-FSL and fine-grained datasets in the 5-way 1-shot setting.

1-shot	Backbone	TTF	ChestX	ISIC	EuroSAT	CropDisease	CUB	Cars	Places	Plantae	Avg.
			ImageNet	Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	
PM [12]	ViT-small	-	22.74 ± 0.40	33.72 ± 0.60	72.94 ± 0.77	81.04 ± 0.85	83.53 ± 0.86	42.10 ± 0.80	71.66 ± 0.88	59.04 ± 0.89	58.35
StyleAdv [†] [86]	ViT-small	-	22.92 ± 0.32	33.05 ± 0.44	72.15 ± 0.65	81.22 ± 0.61	84.01 ± 0.58	40.48 ± 0.57	72.64 ± 0.67	55.52 ± 0.66	57.75
AMT	ViT-small	-	22.39 ± 0.39	33.92 ± 0.58	73.52 ± 0.84	82.04 ± 0.80	84.34 ± 0.83	44.33 ± 0.81	73.78 ± 0.87	59.32 ± 0.94	59.21
PMF [12]	ViT-small	Y	21.73 ± 0.30	30.36 ± 0.36	70.74 ± 0.63	80.79 ± 0.62	78.13 ± 0.66	37.24 ± 0.57	71.11 ± 0.71	53.60 ± 0.66	55.46
StyleAdv-FT [†] [86]	ViT-small	Y	22.92 ± 0.32	33.99 ± 0.46	74.93 ± 0.58	84.11 ± 0.57	84.01 ± 0.58	40.48 ± 0.57	72.64 ± 0.67	55.52 ± 0.66	58.57
AMT-FT	ViT-small	Y	23.23 ± 0.40	33.95 ± 0.63	73.95 ± 0.78	82.04 ± 0.8	84.34 ± 0.83	46.06 ± 0.80	73.83 ± 0.89	59.32 ± 0.94	59.59

Table 3. Few-shot classification robustness against adversarial attacks under distribution shifts.

1-shot	Adv. TTF	In-domain		Out-of-domain										Avg.
		ImageNet	Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO			
PM [12]	-	23.22	7.74	5.37	22.38	25.39	1.11	12.79	24.99	2.23	10.20	13.54		
StyleAdv [86]	-	16.76	15.25	5.95	17.70	25.75	1.43	14.78	30.75	3.07	9.63	14.11		
AMT	-	33.70	42.19	11.72	32.05	32.47	27.45	19.74	41.12	22.79	17.67	28.09		
PMF [12]	Y	23.22	31.77	18.35	22.65	25.39	30.99	23.20	38.93	25.86	23.69	26.41		
AMT-FT	Y	33.70	42.19	20.40	34.92	32.47	37.49	20.10	41.12	32.75	22.70	31.78		

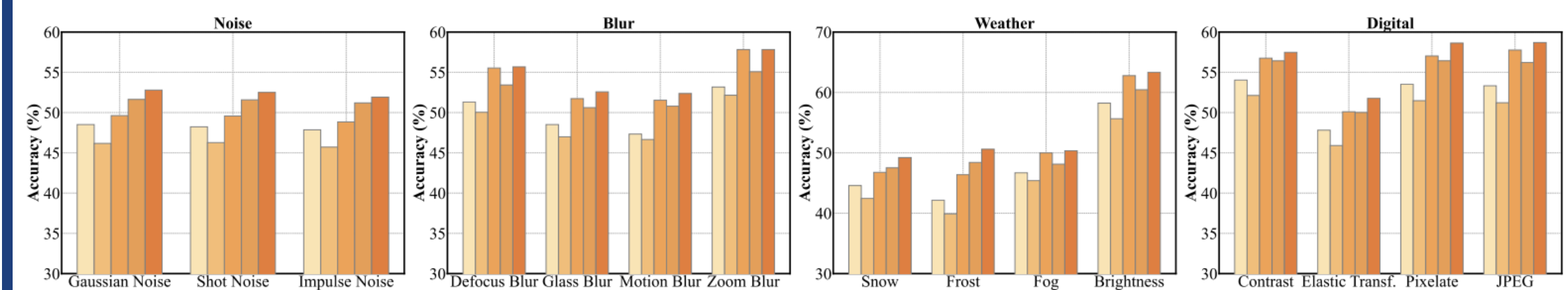


Figure 1. Few-shot classification robustness against image corruptions under distribution shifts.

Analysis

APQ	APSV	RLP	TTM	STr	In-domain		Out-of-domain										Avg.
					INet	Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO			
✗	✗	✗	✗	✗	65.07	59.03	38.13	76.18	61.56	57.29	56.03	80.41	55.17	54.42	60.35		
✓	✗	✗	✗	✗	64.57	62.47	38.53	76.23	60.47	57.97	56.22	81.72	57.04	53.96	60.92		
✓	✗	✓	✓	✗	65.56	63.92	39.74	76.06	61.73	58.64	55.99	80.93	56.96	54.28	61.38		
✓	✗	✓	✓	✓	64.95	70.80	40.55	75.19	60.73	59.66	56.92	83.63	57.66	56.04	62.61		
✓	✓	✓	✓	✗	67.95	62.16	39.13	79.27	61.77	58.75	56.59	79.74	55.45	54.63	61.54		
✓	✓	✓	✓	✓	68.46	65.75	42.63	79.43	63.10	58.23	55.69	78.93	63.67	56.28	63.22		
✓	✓	✓	✓	✓	68.80	71.95	42.90	79.95	62.99	59.62	59.06	85.37	63.78	57.14	65.16		

Achieve SoTA Performance with A Single Adversarially Meta-tuned Model