A. Supplementary Material

A.1. Datasets

CIFAR10 [42] is widely used for image classification. It contains 10 classes with 60,000 images (50,000 for training and 10,000 for testing). Its small resolution (32×32 pixels) and balanced class distribution make it a common benchmark for evaluating adversarial robustness. `CIFAR10` has significantly fewer classes and a lower level of visual complexity compared to `ImageNet21-k`. This enables us to study how fine-tuning on simpler datasets change model robustness inherited from pre-training.

CIFAR100. `CIFAR100` [25] extends `CIFAR10` to 100 classes, each containing 600 images (500 for training, 100 for testing). While it shares the same low-resolution format, `CIFAR100` introduces a more fine-grained classification task. The increased class diversity and hierarchical structure (coarse and fine labels) make it a more complex dataset but still much smaller in scale compared to `ImageNet-21k`.

Caltech256. `Caltech256` [13] comprises 256 classes with 30,607 images, offering significantly more class diversity than CIFAR datasets. It has a minimum of 80 images per class. `Caltech256` contains higher-resolution images with more natural object variations, making it more similar to `ImageNet-21k` in terms of complexity and scale. With this, we can better understand how fine-tuning on a moderately large dataset with varied classes affects robustness.

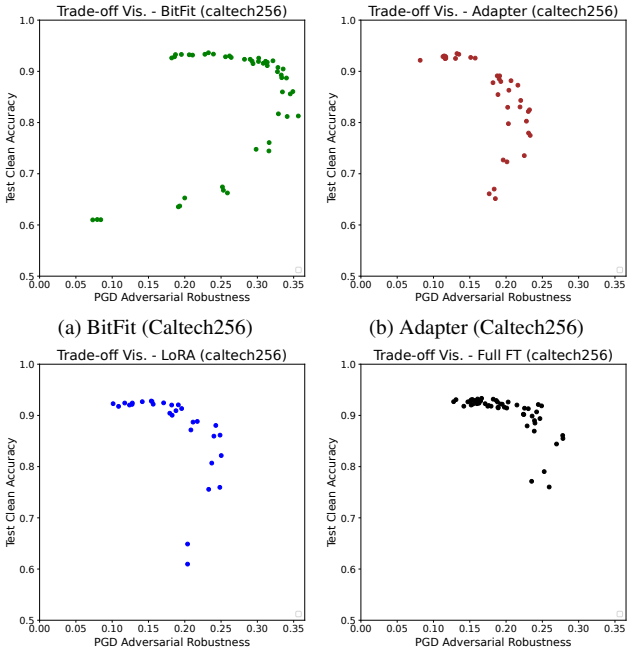
CUB-200-2011. `CUB200` [47] is a fine-grained classification dataset containing 11,788 images across 200 bird species. Unlike broader classification datasets, `CUB200` focuses on a single semantic category (birds), making it an important benchmark for studying adversarial robustness in tasks where pre-trained models are fine-tuned on more specialized, domain-specific knowledge. Since `ImageNet-21k` includes bird species in its taxonomy, this dataset allows us to explore how fine-tuning on a subdomain of the pre-training distribution impacts robustness.

Stanford Dogs. `Stanforddogs` [23] is another fine-grained classification dataset with 22,000 images of 120 dog breeds. Similar to CUB, it provides a challenging adversarial benchmark due to the subtle intra-class variations among breeds. Since `ImageNet-21k` also contains dog breeds, this dataset enables us to investigate whether fine-tuning on a narrower but related distribution affects the robustness inherited from pre-training.

DomainNet. `DomainNet` [41] is a large-scale domain adaptation dataset of 586,575 images, containing six different domains: clip art, info graph, painting, quick draw, real, and sketch. ImageNet-21k primarily contains real-world images, making `DomainNet` an effective benchmark to test how well fine-tuned models generalize when faced with significant distributional changes, particularly when trained on one domain and tested on others.

A.2. Trade-off Space of Fine-Tuning

We present the trade-off space between adversarial robustness and accuracy in Figure 8. This corresponds to the training curves shown in Figure 3.



(a) BitFit (Caltech256) (b) Adapter (Caltech256)

(c) LoRA (Caltech256) (d) Full fine-tuning (Caltech256)

Figure 8. Trade-off visualization for Caltech256. The dots are corresponding to different time stamps during training (from bottom left to upper right to upper left as time goes on).

A.3. Ablation Study

We perform an ablation study on the number and location of trainable parameters and learning rate. We conduct experiment sweeps with final model checkpoints on 1) LoRA rank (r $\in$ [1,20]), 2) Adapter reduction factor ($d \in$ [4,32]), and 3) update location (attention vs. FNN). The results show that robustness does not consistently correlate with the number of trainable parameters or layers updated. This suggests that these factors alone do not explain robustness differences. However, the results all show low adversarial accuracy with small variances. It led us to our design choise: tracking changes over training steps offer deeper insights.

In addition, we also study how different learning rates affect OOD robustness during training. We track OOD robustness while fine-tuning models with varying learning rates—$\{1e-4, 5e-4, 1e-3, 5e-3\}$. As shown in Figure 9, learning rate does impact accuracy and robustness, especially during the early steps, but show consistent trend as claimed across steps and converge in the end.

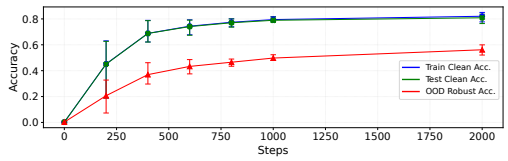


Figure 9. The OOD robustness of models fine-tuned by LoRA with different learning rates on the *"real"* domain and evaluated on 5 other domains.

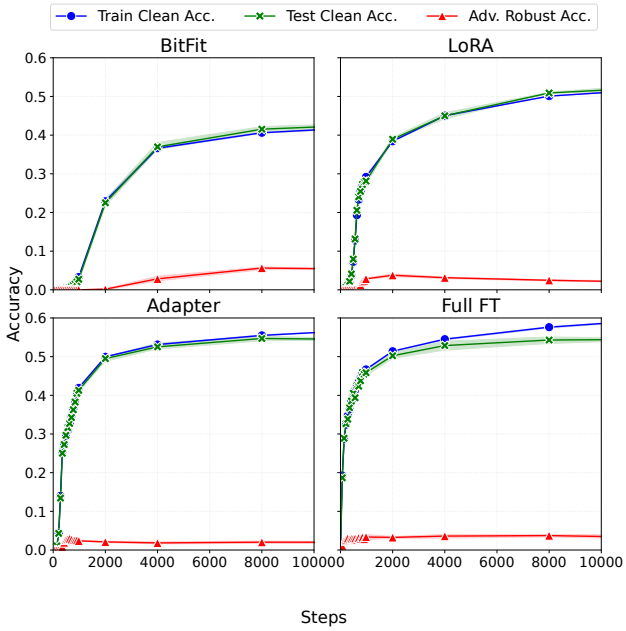


Figure 10. Continuous evaluation of training accuracy (blue), test accuracy (green), and adversarial robustness (red) across back-propagation steps (truncated at 10,000 steps) on Places365 [52].

A.4. Hyperparameters

Grid search is used to find optimal training hyperparameters: base learning rate in $\{1e-4, 1e-5, 3e-5, 5e-5\}$, base weight decay in $\{1-e2, 1-e3\}$, and the adjustment ratio for each fine-tuning strategy in $\{1, 10, 5, 10, 2, 2, 3\}$ (corresponding to the order of the strategies shown in Table 3). These choices are based on previous literature [16, 17, 49] for different fine-tuning methods and downstream datasets. They all have comparable scale of trainable parameters (in percentage), except for full fine-tuning:$\{100, 0.01, 1.19, 0.13, 2.03, 0.07, 0.07\}$. Due to the large size of DomainNet [41], we consistently set the base weight decay to be $1e-2$. The specific optimal training hyperparameters can be found in Table 3 and Table 4.

A.5. Dataset Scale

PEFT is especially relevant in low-data regimes. Our main experiments focus on datasets with 10k to 60k samples, covering diverse degrees of task complexity in terms of number of classes, class separation, and similarity to upstream data. To complement this analysis and assess the effect of data scale, we also include experiments on Places365 [52], a medium-scale dataset with approximately 1.8 million samples. As shown in Figure 10, models achieve relatively low clean accuracy (preaking around 50%) and limited adversarial robustness (less than 5%). Due to this overall weak performance, the trend of adversarial robustness and the differences of robustness across fine-tuning strategies are difficult to distinguish.

A.6. Decomposition of Fine-Tuning

As described in Section 3.2.1, we focus on decomposing PEFT methods along two directions: information location and mechanisms, each having four components. The decomposition for five PEFT methods can be found in Table 5.

Fine-Tuning Methods	Learning Rate / Weight Decay for Adv Exps.				
	CIFAR10	CIFAR100	CUB200	Caltech256	Stanford Dogs
Full Fine-tune	3e-5/1e-3	5e-5/1e-3	5e-5/1e-3	1e-4/1e-3	1e-4/1e-3
Linear Probe	1e-5/1e-3	1e-5/1e-2	5e-6/1e-3	1e-5/1e-3	1e-5/1e-3
LoRA	5e-4/1e-2	5e-4/1e-2	2.5e-4/1e-2	5e-4/1e-3	5e-5/1e-2
BitFit	1e-5/1e-4	1e-5/1e-3	5e-6/1e-4	1e-5/1e-3	1e-5/1e-3
Adapter	1e-4/1e-3	2e-4/1e-2	2e-5/1e-2	2e-4/1e-2	2e-4/1e-3
Compacter	2e-4/1e-3	2e-4/1e-3	1e-4/1e-3	2e-4/1e-3	2e-4/1e-3
$(IA)^3$	1.5e-4/1e-3	3e-4/1e-2	3e-4/1e-3	3e-4/1e-3	1.5e-4/1e-3

Table 3. Strategy configurations with datasets (Adv)

Fine-Tuning Methods	Learning Rate for OOD Exps.					
	Clipart	Infograph	Painting	Quickdraw	Real	Sketch
Full Fine-tune	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
Linear Probe	1e-3	1e-3	1e-3	1e-3	5e-4	1e-3
LoRA	5e-4	2.5e-4	5e-4	2.5e-4	5e-4	5e-4
BitFit	1e-3	1e-3	5e-4	1e-3	5e-4	1e-3
Adapter	2e-4	2e-4	2e-4	1e-4	2e-4	2e-4
Compacter	2e-4	2e-4	2e-4	2e-4	2e-4	2e-4
$(IA)^3$	3e-4	3e-4	3e-4	3e-4	3e-4	3e-4

Table 4. Strategy configurations with datasets (OOD)

PEFT Strategies	**Information Location**				**Mechanism**			
	Attn	FFN	Rep.	Bias	Proj. Layers	Matrix Reparam	Element-wise Mult.	Direct Update
LoRA	●	○	○	○	○	●	○	○
IA3	○	○	●	○	○	○	●	○
Adapter	○	○	●	○	●	○	○	○
Compacter	○	○	●	○	●	●	○	○
BitFit	●	●	○	●	○	○	○	●

Table 5. The space of PEFT strategies in terms of information location and underlying mechanisms