NeurIPS 2024

# TARP-VP: Towards Evaluation of Transferred Adversarial Robustness and Privacy on Label Mapping Visual Prompting Models

***Presenter: Zhen Chen***

Yi Zhang, Fu Wang, Xingyu Zhao, Xiaowei Huang, Wenjie Ruan

UNIVERSITY OF LIVERPOOL

UNIVERSITY OF EXETER

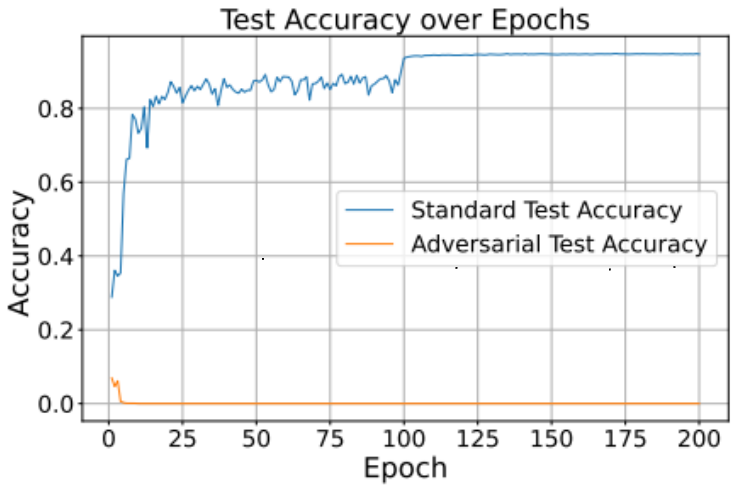
WARWICK
THE UNIVERSITY OF WARWICK

# Outline

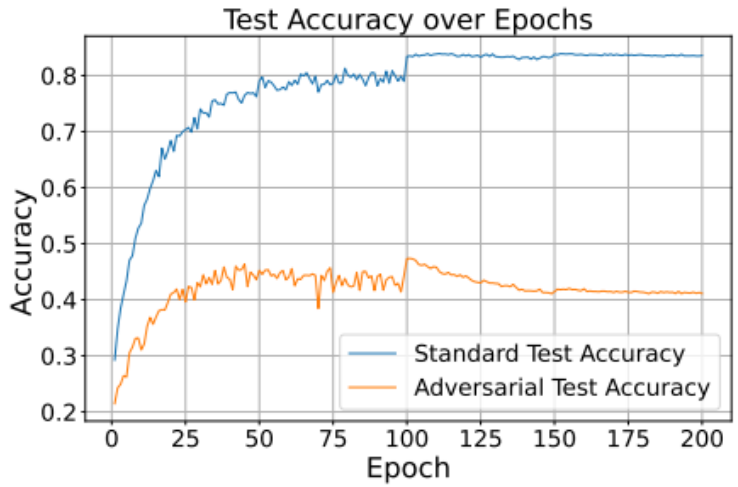
- Motivation
- White-box Adversarial Robustness of LM-VP
- Transferred Adversarial Robustness of LM-VP
- Privacy Evaluation  of LM-VP

# Motivation: Adversarial robustness and privacy trade-off

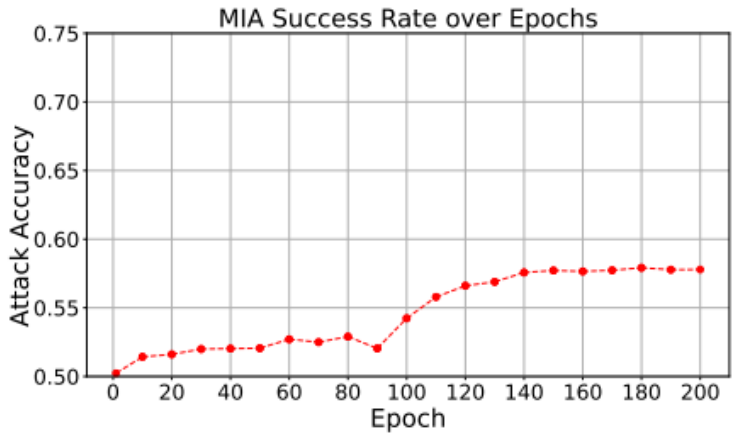
## For a general deep learning model:



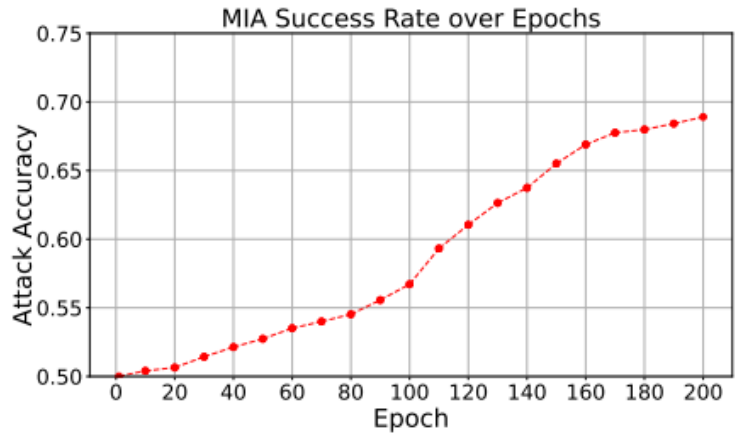

(a) Test Accuracy of Standard Training




(b) Test Accuracy of Adversarial Training




(c) MIA on Standard Training



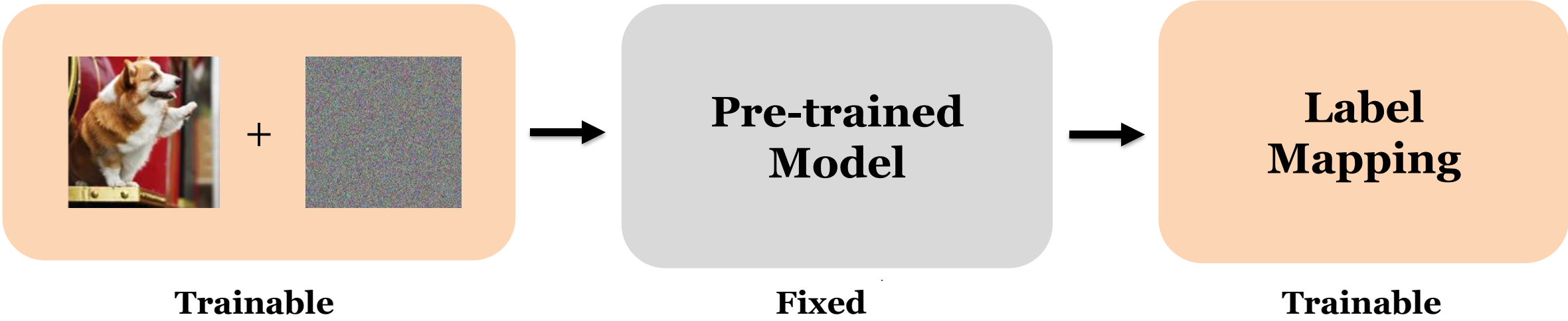

(d) MIA on Adversarial Training

Figure 1: Trade-off between test accuracy and membership inference attacks of standard training and adversarial training along with training on CIFAR-10 with $\ell_\infty$ threat model using ResNet18.

More severe privacy risks on AT:

- Larger generalization error
- Higher sensitivity
- Robust overfitting

# Design of Label Mapping Visual Prompting Models:



# Design of Visual Prompting

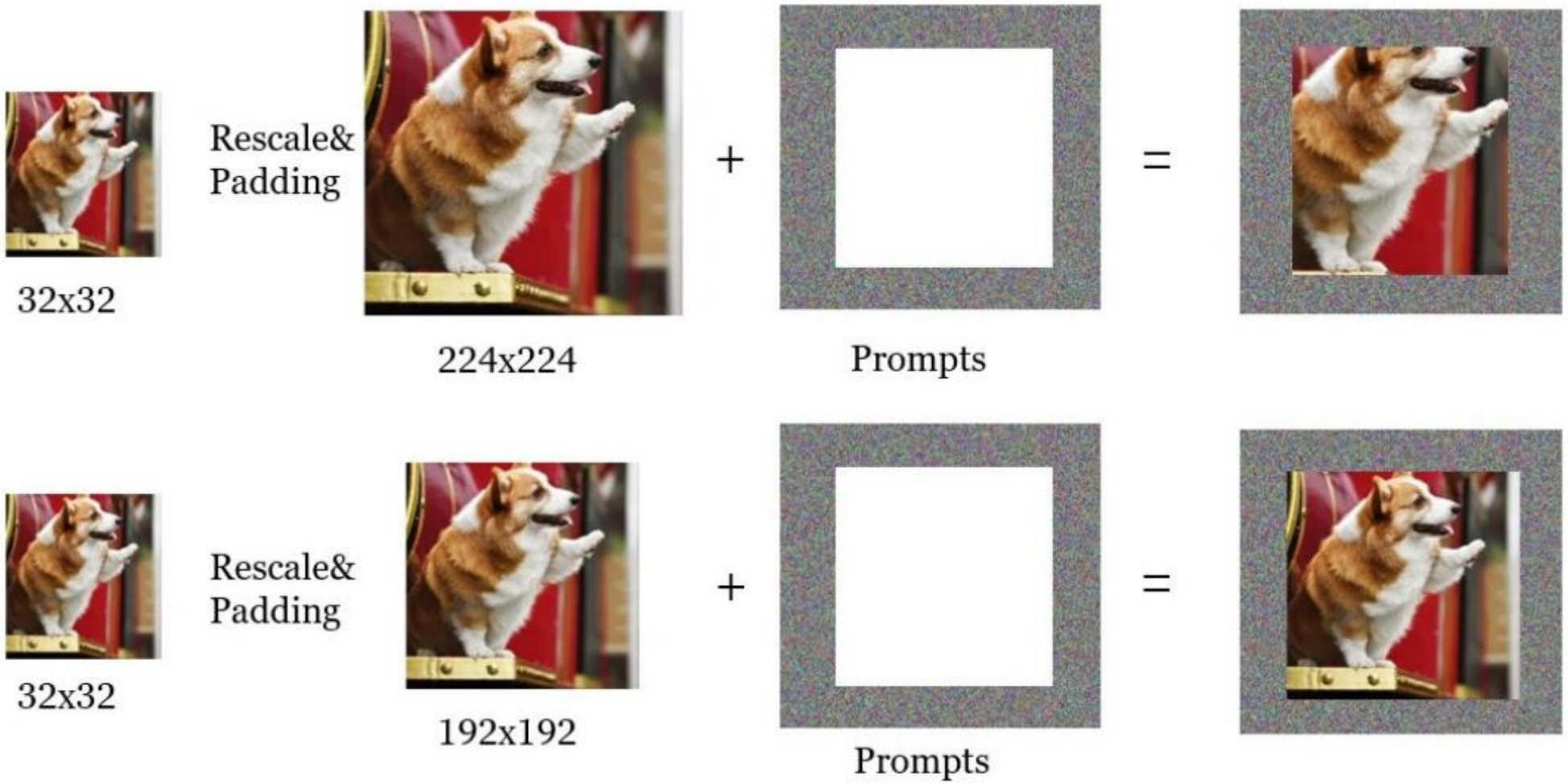
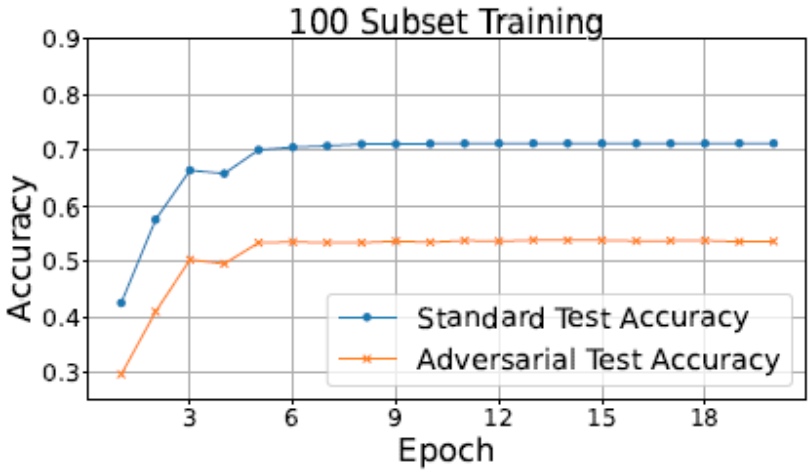


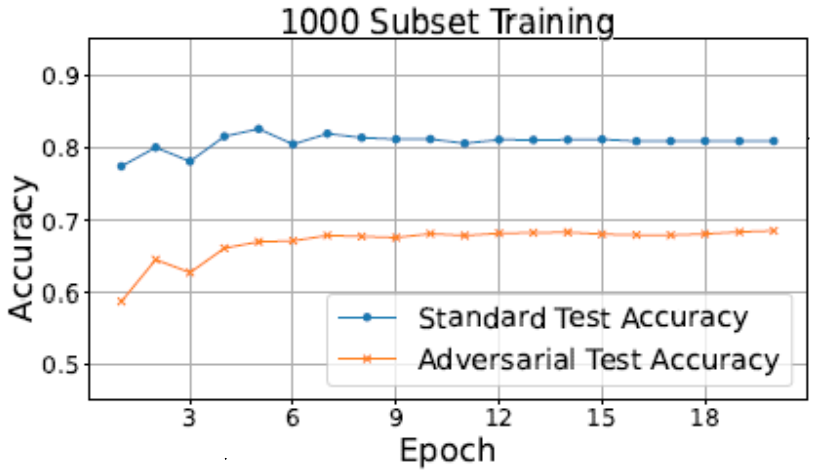

Figure 2: Two ways to add prompts: (1) Top: rescale a target image to the source domain size and replace the edge of the image with prompts; (2) Bottom: rescale a target image to a size smaller than the source domain and add prompts to make it the same size as source domain.

# Characteristic of LM-VP

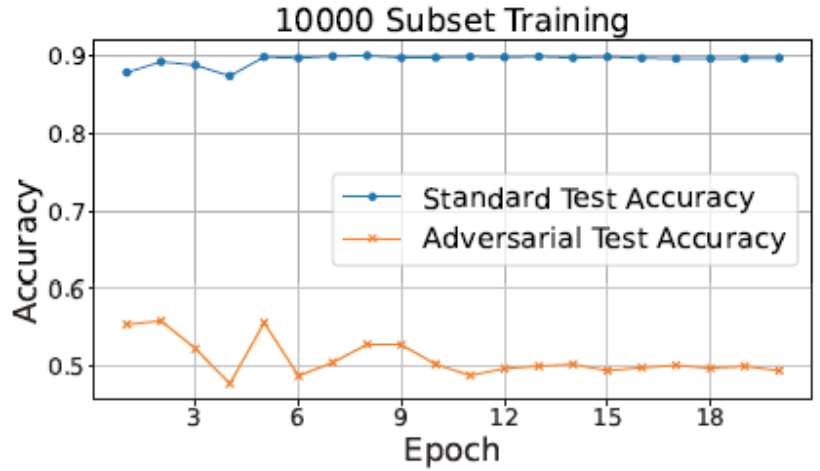
## (1) Insufficient training data in LM-VP



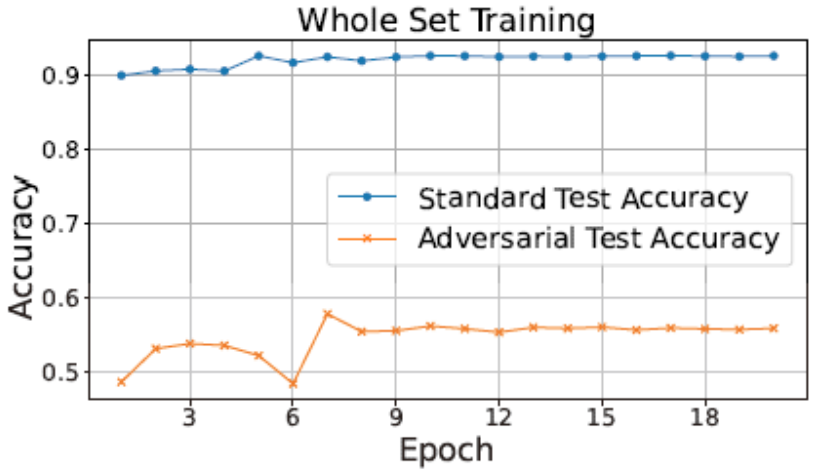

(a) Test Accuracy on Random 100 Subset Training




(b) Test Accuracy on Random 1000 Subset Training




(c) Test Accuracy on Random 10000 Subset Training



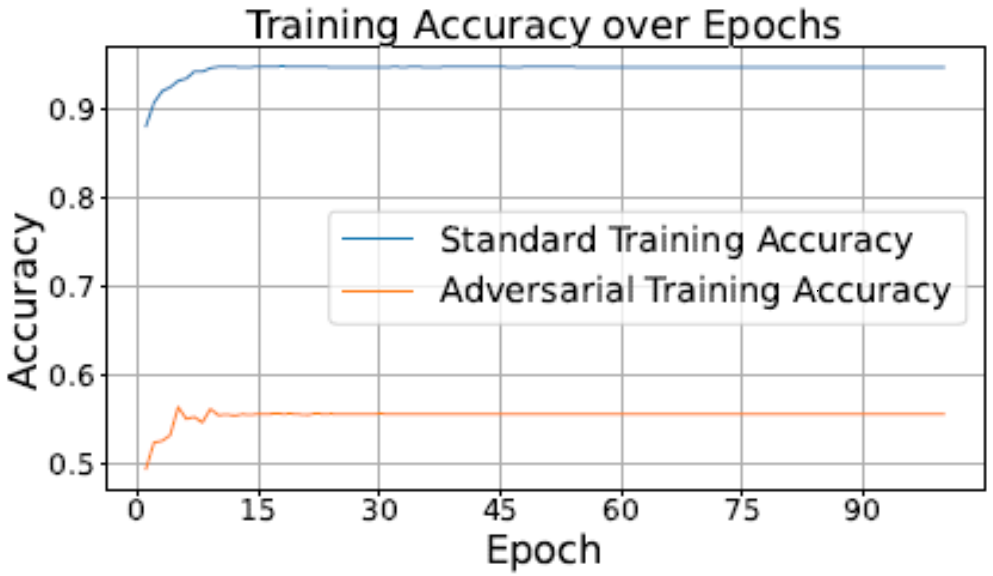

(d) Test Accuracy on Whole Training Set Training set

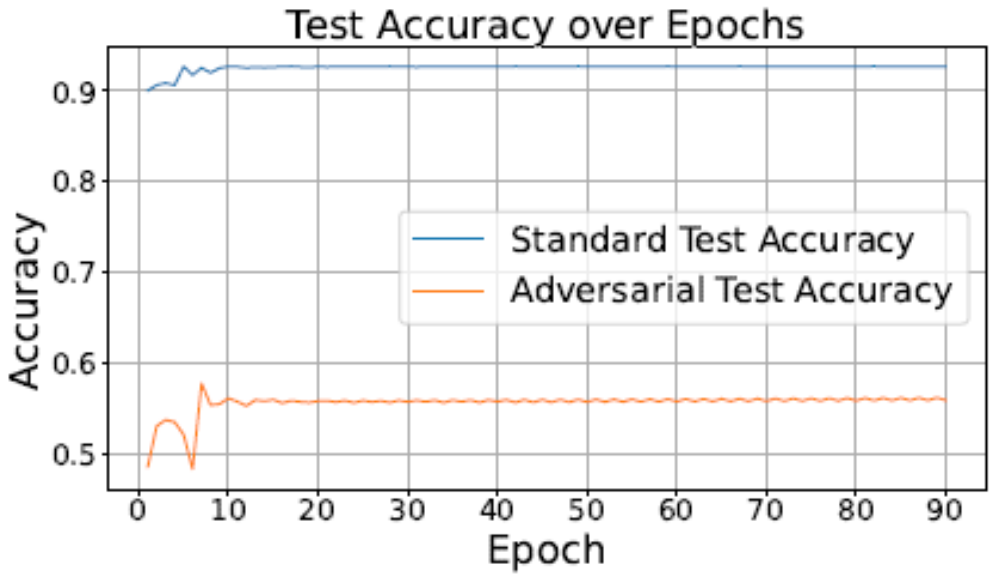
Lower sensitivity of LM-VP:

- Highest transferred adversarial robustness on 1000 subest training
- Similar standard accuracy on 10000 subst training

# Characteristic of LM-VP

## (2) Rapid convergence and minimal generalization error of LM-VP



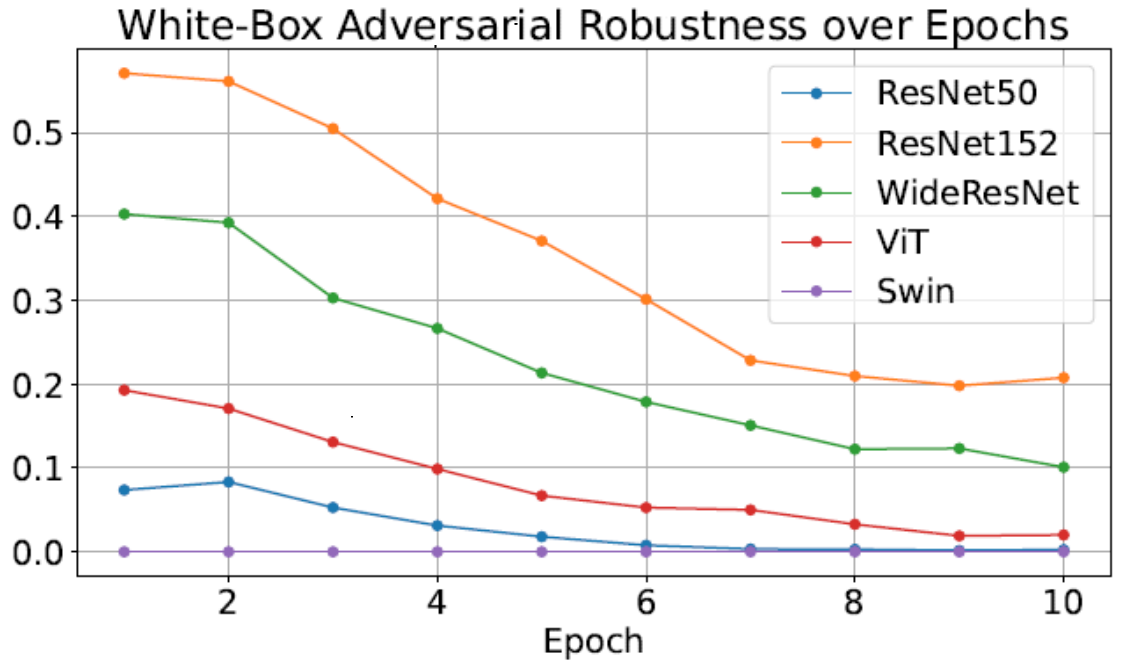

- Quickly achieve a near-optimal performance and then remain steady with continued training
- Minimal generalization error

# White-box Adversarial Robustness of LM-VP

**White-box adversarial robustness of LM-VP is largely influenced by the choice of pre-trained models and there is no clear pattern**

Table 1: Best performance(%) on CIFAR-10 with different pre-trained models in Standard-Trained LM-VP models and Standard AT-Trained LM-VP models under white-box adversarial attacks.

| Pre-trained models | Standard Training | | Adversarial Training | |
|---|---|---|---|---|
| | $\text{Natural}_{te}$ | PGD-20 | $\text{Natural}_{te}$ | PGD-20 |
| **ResNet50** | 80.52 | 8.33 | 23.10 | 0.8 |
| **ResNet152** | 84.76 | 57.09 | 14.24 | 0 |
| **Wideresnet** | 80.91 | 40.29 | 12.15 | 0 |
| **VIT** | 91.50 | 19.28 | 27.78 | 0 |
| **Swin** | 92.00 | 0 | 34.65 | 0 |
| **ConvNext** | 97.97 | 43.22 | 40.69 | 0 |



# Transferred Adversarial Robustness

## Transferred adversarial robustness of standard-trained LM-VP

Table 2: Best performance(%) on CIFAR-10 with different pre-trained models in Standard-Trained LM-VP models under Threat models ResNet18 or WRN-34-10.

| Best Performance on natural examples and adversarial examples | | | | | | | |
|---|---|---|---|---|---|---|---|
| Pre-trained models | Threat models | $\text{Natural}_{tr}$ | $\text{Natural}_{te}$ | $\text{PGD-10}_{tr}$ | PGD-20 | CW-20 | T/E |
| **ResNet50** | ResNet18 | 87.73 | 86.30 | 31.14 | 35.61 | 34.30 | 251s |
| **ResNet152** | | 90.39 | 89.51 | 36.76 | 35.99 | 35.67 | 440s |
| **WRN-50-2** | | 87.77 | 86.78 | 37.73 | 39.76 | 38.90 | 381s |
| **VIT** | | 94.91 | 92.67 | 51.25 | 51.95 | 50.70 | 589s |
| **Swin** | | 94.78 | 92.71 | 56.46 | 57.80 | 57.34 | 1025s |
| **ConvNext** | | 99.33 | 98.28 | **88.70** | **89.11** | **89.37** | 2116s |
| **EVA** | | **99.66** | **98.54** | 86.95 | 87.40 | 87.56 | 2674s |
| **Best Performance on natural examples and adversarial examples** | | | | | | | |
| Pre-trained models | Threat models | $\text{Natural}_{tr}$ | $\text{Natural}_{te}$ | $\text{PGD-10}_{tr}$ | PGD-20 | CW-20 | T/E |
| **ResNet50** | WRN-34-10 | 87.18 | 85.87 | 30.33 | 32.32 | 30.98 | - |
| **ResNet152** | | 89.95 | 89.42 | 37.24 | 37.26 | 37.08 | - |
| **WRN-50-2** | | 87.97 | 87.01 | 38.25 | 41.36 | 39.90 | - |
| **VIT** | | 94.78 | 92.77 | 51.41 | 52.23 | 52.12 | - |
| **Swin** | | 95.08 | 92.8 | 55.23 | 59.20 | 57.54 | - |
| **ConvNext** | | 99.19 | 98.03 | **88.20** | **88.51** | **88.23** | - |
| **EVA** | | **99.64** | **98.45** | 86.21 | 86.98 | 87.24 | - |

# Transferred Adversarial Robustness

## Transferred adversarial robustness of transferred AT-trained LM-VP:

Table 3: Best performance(%) on CIFAR-10 with different pre-trained models in Transfered AT-Trained LM-VP models under Threat model ResNet18.

| Best Performance on natural examples and adversarial examples | | | | | | | |
|---|---|---|---|---|---|---|---|
| Pre-trained models | Threat models | Natural$_{tr}$ | Natural$_{te}$ | PGD-10$_{tr}$ | PGD-20 | CW-20 | T/E |
| **ResNet50** | ResNet18 | 68.84 | 70.37 | 64.10 | 63.01 | 61.78 | 671s |
| **ResNet152** | | 68.83 | 77.08 | 63.39 | 63.95 | 62.92 | 950s |
| **WRN-50-2** | | 69.68 | 70.42 | 62.07 | 62.86 | 60.89 | 875s |
| **VIT** | | 86.23 | 86.64 | 77.49 | 75.34 | 74.87 | 1380s |
| **Swin** | | 89.32 | 89.74 | 80.72 | 79.14 | 77.89 | 2205s |
| **ConvNext** | | 97.79 | 98.02 | 92.61 | 91.63 | 91.02 | 3446s |
| **EVA** | | **98.64** | **98.32** | **93.19** | **92.43** | **91.50** | 4136s |

**Transferred AT significantly enhances the transferred adversarial robustness at the cost of reduced natural accuracy**

# MIA-based Privacy Analysis

**Privacy Analysis of LM-VP models:**
**(1) lower sensitivity of LM-VP models to training data**
**(2) minimal generalization error**
**(3) Prior knowledge embedded in different pre-trained models**

$$MIA(\eta) = \frac{1}{2} \times \left( \frac{\sum_{(x,y)\in D_{\text{train}}} \mathbf{1}\left[f_\theta(x)_y \geq \eta\right]}{|D_{\text{train}}|} + \frac{\sum_{(x,y)\in D_{\text{test}}} \mathbf{1}\left[f_\theta(x)_y < \eta\right]}{|D_{\text{test}}|} \right)$$

$$\eta_{\text{optim}} = \arg\max_{\eta} MIA(\eta)$$

# MIA Evaluation

**Transfer AT improves both transferred adversarial robustness and MIA-based training data privacy**

Table 4: MIA success rate(%) on CIFAR-10 with different pre-trained models in Standard and Transferred AT Trained LM-VP models under Threat model ResNet18.

| Generation Gap and MIA Success Rate on Trained LM-VP Models | | | | |
|---|---|---|---|---|
| Pre-trained models | Standard Training | | Transfered AT | |
| | MIA Nat | MIA Adv | MIA Nat | MIA Adv |
| **ResNet50** | 68.92 | 57.88 | 55.27 | 51.19 |
| **ResNet152** | 75.34 | 56.46 | 62.15 | 50.77 |
| **WRN-50-2** | 62.58 | 50.66 | 50.46 | 50.94 |
| **VIT** | 51.66 | 50.37 | 50.53 | 51.78 |
| **Swin** | 51.75 | 50.53 | 50.23 | 51.63 |
| **ConvNext** | 80.14 | 77.33 | 50.32 | 50.70 |
| **EVA** | 77.46 | 73.35 | 50.32 | 50.67 |

# Conclusion

- Pre-trained models significantly influences the white-box adversarial robustness of LM-VP→hard to draw consistent conclusions
- Transfer AT achieve a good trade-off between transferred adversarial robustness and MIA-based privacy→consistent findings across various pretrained models